Comprehensive Statistical Modeling of Cyber Risks and Analysis of Data Breaches with Applications in Cyber Insurance

by

Meng (Maggie) Sun

M.Sc. (Mathematics), University of Connecticut, 2018
B.Sc. (Actuarial Science), University of International Business and Economics, 2014

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2024

Declaration of Committee

Name: **Meng (Maggie) Sun**

Degree: **Doctor of Philosophy**

Thesis title: **Comprehensive Statistical Modeling of Cyber Risks and Analysis of Data Breaches with Applications in Cyber Insurance**

Committee: **Chair:** Liangliang Wang
Associate Professor, Statistics and Actuarial Science

Yi Lu
Supervisor
Professor, Statistics and Actuarial Science

Himchan Jeong
Committee Member
Assistant Professor, Statistics and Actuarial Science

X. Joan Hu
Examiner
Professor, Statistics and Actuarial Science

Jiandong Ren
External Examiner
Professor, Statistics and Actuarial Science
Western University

Abstract

In the rapidly evolving landscape of cybersecurity, the increased demand for zero trust protection and the intricate management of digital assets give rise to the urgent need for robust cyber risk mitigation strategies. Despite significant investments in information security, the escalating frequency and severity of cyber breaches pose substantial risks to business operations, with potential large-scale economic impacts. This thesis presents a comprehensive analysis of data breaches, employing advanced statistical modeling and estimation techniques. An empirical investigation of the Privacy Rights Clearinghouse (PRC) Data Breach Chronology dataset, including cluster analysis and preliminary data examination, sets the groundwork for subsequent modeling approaches. A Bayesian negative binomial generalized linear mixed model is introduced to capture quarterly variation and heterogeneity in cyber incidents frequency. Further, the thesis proposes a zero-inflated mixture and composite regression model for the loss severity. This model incorporates splicing and finite mixture techniques to address unique features of data breaches, with the parameter estimation facilitated by the expectation-maximization (E-M) algorithm. Building on frequency and severity models, the research introduces aggregate loss modeling approaches, including simple aggregation and MCMC-based methods. These models offer practical strategies for the cyber insurance industry. The impact of various deductibles, limits, and reinsurance practices on loss aggregations is also examined. The findings emphasize the critical importance of accurate cyber risk measurement and prediction for effective risk management and mitigation. By leveraging advanced statistical models, this research contributes to the development of more resilient cybersecurity frameworks and informs strategic decision-making in advancing cyber insurance products.

Keywords: cyber risk aggregation; cyber risk modeling; expectation-maximization (E-M) algorithm; generalized linear mixed model (GLMM); Markov chain Monte Carlo (MCMC); mixture composite regression

Dedication

To my beloved family, with gratitude for their eternal love.

Acknowledgements

I would like to express my deepest gratitude to the professor who brought me in as the first actuarial science Ph.D. candidate: Professor Yi Lu. She is gifted in Risk theory, Stochastic modeling, and Statistical application. I am honored to have her as my supervisor for my Ph.D.. Her academic professionalism and thorough knowledge lit up my research pathway ahead. She is caring and easy-going and is always willing to offer assistance when I run into obstacles. She backed me up as much as possible, not only financially but also academically; I was so lucky that I could go both far and deep on the topic interests me. Her unwavering support and endless patience both to my research and life enable me to accomplish the ending phase of my highest degree. I am forever in debt to her dedicating time and energy to my success, for her countless hours having discussions, revising manuscripts, encourage me to attend conferences and make presentation, believing in my research capability and never giving up on me.

My family is unconditionally and persistently supportive for all my ideas and decisions. Its infinite love and endless encouragement enabled me to concur anxiety and depression during the most frustrating period that I experienced. Two little cutey cats, Husky and Anthony, showed up in different phases of my life, accompanied me and completed me. Wish you two all happily ever after.

I appreciate the academic environment and opportunities that the Department of Statistical and Actuarial Science offered to our students. My special thanks go to Professor Richard Lockhart, for his vivid lecture, versatile knowledge and especially for meaningful advice in statistical learning when I was preparing for comprehensive exam with a non-statistical background. I am also thankful to Professor Boxin Tang for willingness to spend time explaining course materials in detail and encouraging me to perfect my English speaking. Fortunately I had meaningful experiences working as teaching assistants instructed by Professor Yi Lu, Professor Chi-Liang (Cary) Tsai.

I would like to thank my thesis defense chair, Dr. Liangliang Wang, for her invaluable guidance and leadership throughout the defense process. I extend my heartfelt thanks to my commitee member, Dr. Himchan Jeong. Your critical insights and suggestions have significantly improved the quality of this thesis. Your dedication to my success and your willingness to challenge and inspire me have been crucial in my academic journey. I am profoundly grateful to my internal examiner, Dr. X. Joan Hu, for her meticulous review

and constructive feedback. Your thorough understanding of the subject matter and your attention to detail have ensured that this thesis meets the highest academic standards. Your support and encouragement throughout the process have been deeply appreciated. To my external examiner, Dr. Jiandong Ren, I owe a great debt of gratitude for your time, effort, and expertise. Your objective evaluation and insightful comments have provided a valuable external perspective, greatly enriching this thesis. Your commitment to excellence and your thoughtful critique have been essential in refining and enhancing my work. Thank you all for your contributions to this thesis.

I want to thank my fellow graduate students in the department for their support and friendship. Special thank go to my buddy, Dr. Yuping Yang, for his guidance in fundamental statistics learning. I particularly appreciate his patience and encouragement in tutoring me course materials and programming coding. I also would like to thank my friend Dr. Lulu Guo for her company through the past 5 years in both my personal and research life.

I would like to express my sincere gratitude to Simon Fraser University for providing me the financial support through the Graduate Dean Entrance Scholarship throughout my years of doctoral studies. This support was instrumental in enabling me to dedicate myself fully to my research and academic endeavors.

I could never have made this far without the help and support from my family, professors, and friends. This accomplishment would not have been possible without your guidance and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cyber risks never stand still: a secure remote and hybrid work environment continues to drive investment so demand for zero trust protection for remote workers and organizations is increasing, and at the same time, multicloud environments increase security risks and the complexity of managing them. With border-less network thoroughly covered nearly every terminal in the world, it is crucial to maintain security on digital assets/property and identify vulnerable data residencies in time. Industries, companies and organizations have been increasingly suffered by cyber breaches, which have posed serious risks to their business operations over last decades. According to Gartner (2022), spending on information security and risk management products and services is forecast to grow 11.3% to reach more than $188.3 billion in 2023. Given the potential economic impact of a successful large-scale cyber attack, cybersecurity risks remain the second-most important emerging issue highlighted by risk experts (AXA, 2019). Beyond traditional technologies, a larger set of risks at the intersection of technology and society is rapidly emerging. Artificial intelligence (AI) and big data have led emerging technologies, transforming economic and social structures. The full-scale implications of cyber threats are yet to be experienced, especially since technology is rapidly evolving. This emerging risk becomes a big challenge for the impacted industries and organizations to mitigate and manage the cyber related risks they face and meanwhile for the insurance companies to manage the cyber insurance risk that is transferred to them.

1.1 Background

The increasing trend observed in data breach frequency and associated costs raises the importance of utilizing cyber insurance for business and organizations to protect themselves against data breach losses/liabilities. Cryptomining, phishing, trojan and ransomware have become the biggest cyber threats to businesses lately. A recent industry survey (Rudolph, 2022) indicates that cyber/networks have been listed as number one or two among the top five notable emerging risks in their 2018-2021 surveys. COVID-19 has compelled businesses to establish remote workforce and utilize cloud-based platforms. Due to the pandemic, remote work and digital transformation further increased the average total cost of a data breach incident. The Federal Bureau of Investigation (FBI) reports a 300% increase in reported cyber crime activities since the COVID-19 pandemic began. According to IBM (2020), data breach costs increased from \$3.86 to \$4.24 million per incident on average in 2021, the highest cost in the 17-year history of the report. Cybersecurity & Infrastructure Security Agency (CISA, 2022) states that recent cyber attack activities by Russian have included destructive malware and ransomware operations, which changes the cyber threat landscape and leads to global supply constraints. Attackers stole \$121,000 in bitcoin through nearly 300 transactions due to a Twitter breach that affected 130 accounts (Leswing, 2020) resulted in attackers swindling. A ransomware attack paralyzed at least 200 U.S. companies via Kaseya, a globally used software supplier on July 3, 2021 (BBC News, 2021). It was a colossal and devastating supply chain attack and has the potential to spread to any size or scale business through cloud-service providers. A security breach disclosed by Marriott compromised the data of more than 5.2 million hotel guests (Marriott, 2020). The Equifax Data Breach, occurred in 2017 at the American credit bureau Equifax, costed up to \$425 million in total; it affected 147.9 million consumers (Equifax, 2017). All we have seen shows that cyber attacks continue to remain a top threat in future armed conflicts, energy shortages and supplement traditional forms of warfare. The increasing number of large-scale and also widely publicized security breaches suggests that both the number of security/data breaches and their severity is increasing.

Regulations and best practices in cyber security hygiene and risk management are changing due to the rapid increases of frequency and severity of cyberattacks. Several federal legislation (e.g., Data Security and Breach Notification Act (2015) and Data Accountability and Trust Act (2019)) have been introduced in the U.S. to enhance the cyber security and data protection. Prominent legislation like the European Union's 2018 General Data Protection Regulation (GDPR)[1], California's 2020 California Consumer Privacy Act (CCPA)[2]

[1] General Data Protection Regulation: https://gdpr.eu/

[2] California Consumer Privacy Act: https://oag.ca.gov/privacy/ccpa

and Illinois's 2018 Biometric Information Privacy Act (BIPA)[3] have been passed to enforce severe consequences. To collect, store, process and transfer consumer data, these regulations all have one thing in common: they require organizations to adhere to specific standards. The FBI set up an Internet Crime Complaint Center (IC3) (FBI, 2000) in 2000 with a trustworthy source for information on cyber criminal activities to combat through criminal and cyber investigative work. In 2020, IC3 received a total of 791,790 cyber crime records from American public with reported losses exceeding \$4.1 billion, which is a 69% increase in total complaints and about 20% increase in loss amount from 2019. Over the years from 2016 to 2020, IC3 received over two million complaints, reporting nearly \$13.3 billion (FBI, 2000) loss. Those complaints address a wide array of Internet scams affecting victims across the globe.

In addition to reducing vulnerable exposure and increasing technology defence investment, cyber insurance is a fundamental and wildly applicable tool for organizations to maintain their enterprise solvency in light of the rise in cybersecurity threats. Cyber insurance is a type of insurance intended to product against the financial costs associated with the failure or compromise of an organization's information system (Michael A. Bean, 2020). Cyber events incidents include a hacking attack by an external party or malware infection, fraud involving debit and credit cards, and the unintentional disclosure of electronic records due to human error. Cyber insurance is emerging as an important tool to protect organizations against future cyber breach losses and its institutional pillars are progressively evolving and reinforcing one another (Kshetri, 2020). In the 1990s, the earliest form of cyber liability policies were generated to cover online media or errors in data processing, they then evolved to cover unauthorized access, network security, data loss and virus-related claims in the 2000s (ColonyWest, 2023). By analyzing the U.S. cyber insurance market, Xie et al. (2020) find that professional surplus insurers and insurers with surplus insurer affiliation demonstrate a competitive advantage in cyber insurance participation.

According to the Council of Insurance Agents & Brokers' Commercial Property/Casualty Market Survey (CIAB, 2021), average cyber renewal premium rate increases have decelerated, with a 15% sequential-quarter increase in 4Q22 down considerably from a record 34% increase in 4Q21. Fitch (2023) estimates that industry statutory direct written premiums for cyber coverage in standalone and package policies increased by over 50% in 2022 to \$7.2 billion, following 73% premium growth in the prior year. Standalone cyber coverage, representing approximately 70% of industry premiums, increased by 62% in 2022. Rapid recent premium growth and a reduction in claims experience in 2022 led to a strong recovery in results for the U.S. cyber insurance line following two consecutive years of more elevated loss ratios. A significant improvement of the direct incurred loss and defense and cost containment (DCC) expenses ratio drops from 68% to 43% in year 2022 for standalone

[3] Illinois Biometric Information Privacy Act: https://www.ilga.gov/legislation/ilcs

cyber coverage. The NAIC report (NAIC, 2020) points out that changes in cyber insurance loss ratios are not driven by premium growth but by claim frequency and severity growth, implying the significance of cyber insurance policy designs. Most cyber insurance providers a core set of coverages and various supplement coverages. With the cyber risk insurance market is at an inflection point, it provides an opportunity to embrace a paradigm shift. To safeguard its profitability, the cyber insurance market took four deliberate measures to combat rising loss ratios: (Farley, 2022): cyber premiums increased across the board, regardless of the industry sector or organization size; many carriers imposed sub-limits and coinsurance provisions specific to ransomware claims; carriers wanted to limit their exposure by limiting capacity; and almost all carries requested more information regarding data security control efforts.

1.2 Motivation

As it can be seen from the facts and challenges presented in the last section that industries, societies, governments as well as the insurance companies face, it is crucial and urgent for insurance companies develop new and/or strengthen existing insurance products to help mitigate risks caused by cyber attacks. Given the difficulties in quantifying this emerging and evolving risk for pricing and risk management due to its discovered and hidden features, new methodologies and techniques, in addition to standard/traditional actuarial methods, need to be developed.

Modeling cyber related risks has become an increasingly important research topic in many disciplines. Eling (2020) presents a comprehensive review of academic literature on cyber risk and cyber insurance in actuarial science and business related fields including economics, finance, risk management and insurance. Here, we briefly review recent research in actuarial science literature on the modeling and analyzing data breach related cyber risks. Maillart and Sornette (2010) reveal an explosive growth in data breach incidents up to July 2006 and a stable rate thereafter. Wheatley et al. (2016) focus on the so called extreme risk of personal data breaches by detecting and modeling the maximum breach sizes and show that the effect of both the frequency and severity of breaches scale is unearthed. Edwards et al. (2016) find that daily frequency of breaches can be well described by a negative binomial distribution. Eling and Loperfido (2017) implement frequency analyses on different levels of breach types and entities through multidimensional scaling and multiple factor analysis for contingency tables, while Eling and Jung (2018) extend former work by implementing pair copula construction (PCC) and Gaussian copula to deal with asymmetric dependence of monthly losses (total number of records breached) in two cross-sectional settings. Fahrenwaldt et al. (2018) develop a mathematical (network) model of insured losses incurred from infectious cyber threats and introduce a new polynomial approximation of claims together with a mean-field approach that allows computing aggregate expected losses and pricing cyber insurance products. Jevtić and Lanchier (2020) propose a structural model of aggregate cyber loss distribution for small and medium-sized enterprises under the assumption of a tree-based local area network (LAN) topology. Schnell (2020) shows that the frequently used actuarial dependence models, such as copulas, and frequency distributions, such as Poisson distribution, would underestimate the strength of dependence.

What we aim to achieve is to generate complex loss prediction models with high predictive accuracy by investigating historical incurred breach incidents and related features. We adopt frequency-severity loss modeling approach, and take the discrete variable for loss frequencies (the number of cyber data breach incidents) and the continuous variable for loss amounts (the recorded number of data breached and the associate dollar amount of loss caused). Those improvements are made to increase prediction accuracy and reduce the bias of prediction models. It is crucial and urgent for researchers and practitioners to identify and model cyber risks such as data breaches so as to help insurance companies examine, pricing and manage their cyber related insurance risks, for which data plays an important role.

Our research is data-driven based on Privacy Rights Clearinghouse (PRC) Data Breach Chronology[4] database which contains cyber breach incidents between years 2001 and 2022. The dataset we examine is from Privacy Rights Clearinghouse (PRC) (PRC, 2019). It is primarily grant-supported and serves individuals in the United States. This repository keeps records of data breaches that expose individuals to identity theft as well as breaches that qualify for disclosure under the state laws. Chronology includes the type of breaches, type of organization, name of company and its physical location, date of incidents and number of records breached. It is the largest and most extensive dataset that is publicly available and has been investigated by several research papers from various perspectives.

1.3 Literature Review

In this section, we provide a literature review on three areas, the generalized linear mixed model (GLMM) for event frequencies, the mixture components spliced regression distribution for loss amounts, and the aggregation of losses or risks. The three parts of review correspond to the models developed/studied and presented in Chapters 3-5 of this thesis, respectively.

1.3.1 Generalized Linear Mixed Model

Generalized Linear Mixed Model (GLMM) is a framework that encompasses statistical approaches to overdispersion, correlated errors, shrinkage estimation, and smoothing of regression relationships. Given an unobserved vector of random effects, observations are as-

[4] Data Breach Chronology, https://privacyrights.org/data-breaches

sumed to be conditionally independent with means that depend on the linear predictor and conditional variances (Breslow and Clayton, 1993). McCulloch (1997) describes maximum likelihood algorithms for GLMMs by constructing a Monte Carlo version of the expectation–maximization (EM) algorithm, proposing a Monte Carlo Newton-Raphson algorithm, and evaluating and improving the use of importance sampling ideas. Clayton (1996) brings up Bayesian analysis using GLMMs to tailor statistical methods not to ill-fitting procedures solely for reasons of computational tractability. Bolker et al. (2009) utilize GLMMs to provide a more flexible approach for analyzing non-normal data such as counts or proportions that often defy classical statistical procedures when random effects are presented. GLMM is one of the most useful structures in modern statistics, allowing many complications to be handled within linear model framework (McCulloch, 2006).

Generally, a generalized regression model is used to describe within-group heterogeneity of observations, and a sampling model is used to describe the group specific regression parameters. A GLMM can handle those issues by not only accommodating non-normally distributed responses and specifying a non-linear link function between response mean and regressors but also allowing group specific correlation in data. In actuarial science literature, Blough et al. (1999) present a GLMM approach to modeling the second part of two-part models for medical expenses utilizing extensions of the generalized linear model; the maximum likelihood method, the generalized quasi-likelihood and the extended quasi-likelihood are discussed. Scurrah et al. (2000) use Bayesian inference with Gibbs sampling to fit GLMMs for right-censored survival times in nuclear and extended families for genetic research which includes age-at-onset and age-at-death data and a variety of other censored traits. Antonio and Beirlant (2007) use the GLMMs in actuarial statistics for the modeling of longitudinal data and discuss the model estimation and inference under the Bayesian framework, in order to offer a solution facing with the fact that actuaries very often have repeated measurements of longitudinal data. Recently, Jeong et al. (2021) quantify and explain the contribution of the variability of claims among policyholders through the use of random effects using generalized linear mixed models, where the aggregate loss is expressed as a product of the number of claims (frequency) and the average claim amount (severity) knowing the frequency; they further calibrate the model using a portfolio of auto insurance contracts from a Singapore insurer. The GLMM has also been used in studying the credibility models; see, for example, Antonio and Beirlant (2007) and Garrido and Zhou (2009).

We present briefly in Section 1.2 the PRC dataset that drives our study. Below are notable studies based on this dataset. Edwards et al. (2016) develop Bayesian generalized linear models to investigate trends in data breaches, and find that the size of data breaches is well modeled by the log-normal family of distributions and that the daily frequency of breaches is described by a negative binomial distribution. Eling and Loperfido (2017) investigate this dataset under the statistical and actuarial framework; multidimensional scaling and goodness-of-fit tests are used to analyze the distribution of data breach information, showing that different types of data breaches need to be modeled as distinct risk categories. Eling and Jung (2018) propose the copula modeling by separating the dependence into pairwise non-zero losses and zero loss arrivals for modeling cross-sectional dependence of data breach losses; copula models are implemented to identify the dependence structure between monthly loss events (frequency and severity). Carfora and Orlando (2019) propose an estimation of value at risk (VaR) and tail value at risk (TVaR) on aggregate losses in dealing with operational risks, and cyber risks in particular. Xu et al. (2018) model hacking breach incident inter-arrival times and breach sizes by stochastic processes and propose data-driven time series approaches to model the complex patterns exhibited by the financial data, showing that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage. Recently, Farkas et al. (2021) present a method for cyber claim analysis based on regression trees to identify criteria for claim classification and evaluation, and Bessy-Roland et al. (2021) propose a multivariate Hawkes framework for modeling and predicting cyber attacks frequency and demonstrate the ability of Hawkes models to capture self-excitation and interactions of data breaches depending on their type and targets.

Several studies in literature discuss Bayesian inference for GLMMs. Fong et al. (2010) conclude that Bayesian inference is now practically feasible for GLMMs and provide an attractive alternative to likelihood-based approaches such as penalized quasi-likelihood. Yau et al. (2003) consider an application of the GLMM approach to the analysis of repeated insurance claim frequency data in which a conditionally fixed random effect vector is incorporated explicitly into the linear predictor to model the inherent correlation, and a motor insurance data set is used as the basis for simulation to demonstrate the advantages of the method. While most studies on modeling cyber risk related dependencies in literature are geared toward cross-sectional dependence using copulas (see, for example, Eling and Jung (2018) and Schnell (2020), and references therein), our approach models the dependence between the frequency and severity under the widely known generalized linear framework, which excels in interpreting the directional effect of features, along with the GLMM that deals with hierarchical effects and dependent variables using general design matrices (McCulloch and Searle, 2004). The Bayesian approach and Markov chain Monte Carlo (MCMC) method are utilized to obtain posterior distributions of parameters of interest. Specifically, our hierarchical structure of Bayesian NB-GLMM requires Metropolis-Gibbs (M-G) sampling schemes working on regression mean related parameters, and conditional maximum likelihood estimates of the dispersion parameter.

1.3.2 Mixture Components Spliced Regression Distribution

A growing number of disciplines are exploring and analyzing Cyber related risks. However, the actuarial cyber risk management is hindered by the need for mature predictive analysis approaches for quantifying and predicting risk severity. We review quantitative research works in actuarial science and describe several research works that focus on loss severity modeling and predictive analysis. Malavasi et al. (2021) combine regression models based on the class of Generalized Additive Models for Location Shape and Scale (GAMLSS), which permits parameters in both the severity and frequency distributions, and a class of ordinal regressions. Giudici and Raffinetti (2021) work on ordinal cyber risk data and propose a rank-based statistical model aimed at predicting the severity levels of cyber risks. Sheehan et al. (2021) propose a conceptual cyber risk classification and assessment framework, designed to demonstrate the significance of proactive and reactive barriers in reducing companies' exposure to cyber risk and quantify the risk. Eling and Jung (2022) measure the size of risk based on the estimation results and show a large degree of heterogeneity across financial firms. Sun et al. (2021) model hacking data breaches frequency using a hurdle Poisson model and severity using a non-parametric generalized Pareto distribution (GPD). Farkas et al. (2021) particularly focus mainly on severe claims by combining a generalized Pareto modelling and a regression tree approach for severity analysis. Most of these methods pay special attention to large claims with heavy tail distributions.

Traditional actuarial modelling techniques for heavy-tailed insurance loss data concentrate on simple models based on a single parametric distribution that adapts the tail well, such as generalized linear models (GLMs), regression models and quantile regression (McNeil, 1997). Buch-Larsen et al. (2005) propose an estimator obtained by transforming the data set with a modification of the Champernowne cdf and then estimating the density of the transformed data by use of the classical kernel density estimator. Charpentier and Oulidi (2010) suggest several nonparametric quantile estimators based on Beta kernel and apply to transformed data by the generalized Champernowne distribution initially fitted to the data. Ahn et al. (2012) study the class of Log phase-type (LogPH) distributions as a parametric alternative in fitting heavy tailed data, which exhibits several advantages over other parametric alternatives. The fact that these techniques are based on a single distribution, which may not be applicable when the behaviour of the tail is inconsistent with the behaviour of the entire loss distribution, highlights a significant limitation in their ability to accurately model and predict extreme loss events. It is well known that the actuarial loss distribution is strongly skewed with heavy tails and consists of small, medium and large claims that are difficult to fit with a single parametric distribution. The Extreme Value Theory (EVT) approach, which employs GPD to model excesses over a high threshold (Allen et al., 2013; Park and Kim, 2016), gains popularity when dealing with heavy-tailed large loss amounts data.

However, above literature fail to capture the characteristics across the entire loss distribution range making them unsuitable for use as a global fit distribution (Beirlant et al., 2004). In order to model the complete loss distribution, it is frequently necessary to obtain a global fit for the distribution of losses by splicing (Klugman et al., 2012) several distributions in order to model the complete loss distribution. Several actuarial works proposed splicing models for the application of risk measures. For financial risk analysis, Reynkens et al. (2017) suggest a splicing model with a mixed Erlang (ME) distribution for the body and a Pareto distribution for the tail. Gan and Valdez (2018) suggest a three-component spliced regression model for fitting insurance loss data and demonstrate that spliced results outperform the Tweedie loss model regarding tail fitting and prediction accuracy. Lim et al. (2011) develop a method for organizing all possible sequence motifs into clusters based on the genomic profile of their positional distribution around splice sites. Poudyal (2021) propose and develop a method of truncated moments (MTuM) and generalize it for different scenarios of loss control mechanism. Blostein and Miljkovic (2019) develop a statistical methodology for fitting left-truncated loss data by using the G-component finite mixture model with any combination of Gamma, Lognormal, and Weibull distributions.

The risk portfolio typically contains unobserved heterogeneity in terms of claim severity, such as workers' compensation and cyber risk data. Given this reality, researchers typically employ a mixture approach to capture the multi-modality of the observed loss distribution. Hathaway (1986) illuminates the relationship of EM for mixture problems by certain clustering techniques and explains global convergence properties of the algorithm without direct reference to an incomplete data framework. Diebolt and Robert (1994) present approximation methods which evaluate the posterior distribution and Bayes estimators by Gibbs sampling, relying on the missing data structure of the mixture model. Everitt (2013) indicates the practical details of fitting finite mixture distributions to sample data. Arcidiacono and Jones (2003) develop a broad class of estimators for mixture models and show this sequential estimator can generate large computational savings with little loss of efficiency. Tzougas et al. (2014) design an optimal Bonus-Malus system in automobile insurance using finite mixture models. Sattayatham and Talangtam (2012) model an actual motor insurance claim dataset using a mixture Lognormal distribution. Bermúdez and Karlis (2012) apply a finite mixture of bivariate Poisson regression models to an automobile insurance claims dataset and insurance ratemaking. Bernardi et al. (2012) propose a finite mixture of skew-normal distributions that better describes insurance data. Miljkovic and Grün (2016) suggest a different method for modelling mixture data with heavy tails and skewness in insurance loss distribution that exhibit multi-modality. Gui et al. (2018) propose an Erlang loss model using a generalized expectation-maximization (GEM) and clustered method of moments (CMM) algorithm to fit insurance loss data and calculate quantities of interest for insurance risk mixture management. Followed by Fung et al. (2019b) propose a class of logit-weighted reduced mixtures of experts (LRMoE) models for multivariate claim frequencies or severity distributions and perform the estimation and application to correlated claim frequencies (Fung et al., 2019a). Recently, Fung et al. (2024) develop a novel class of soft splicing models that bridges the gap between pre-existing methods for handling heavy-tail phenomenon and multi-modality of a claim severity distribution.

1.3.3 Aggregation of Cyber Breaches Risk

Aggregate loss models are used by insurers to segment risk groups, set pure premium, set reserving fund and optimize extreme loss management. The aggregate loss is the summation of all random losses that happened to exposure units in a period of a Property and Casualty (P&C) insurance portfolio. Common practice include evaluating stability of selected variables, grouped levels and interactions using test data, and evaluating model lift and stability of indications. Data sets from data warehouse and third party vendors are used to determine final parameters and indicated relativities. In practical, two approaches are commonly used in estimating the capital under the Loss Distribution Approach (LDA). First is the pure premium or loss cost method, which focuses on the loss ratio—the losses incurred per unit of pure premium—and automatically produces relativities. It is very commonly used for premium estimations in P&C insurance as it requires only one model to build and maintain, allows only a binary choice for the inclusion of a variable and implements offsets easily. The model normally assumes a compound Poisson distribution with gamma claim sizes, and generalize linear models (GLMs) are used to estimate the mean aggregate loss. Since the Tweedie distribution allows to parameterise the compound Poisson-gamma distribution as a member of the exponential dispersion family, it enables the estimation of mean aggregate loss using GLMs directly (Quijano Xacur and Garrido, 2015). In literature, some improvements and innovative contributions have been made to this approach. Araiza Iturria et al. (2021) propose a stochastic model which integrates a double generalized linear model representing both the mean and dispersion of loss ratio distributions, an auto-correlation structure between loss ratios of development lags and a copula-based regression of risks model diving the dependence across various business lines. Denuit et al. (2021) propose auto-calibration in Tweedie-dominance premium calculation model to correct for bias by adding an extra local GLM step to the analysis, in order to minimize Tweedie deviance. Clark (2022) suggests the quasi-Negative Binomial (QNB) as an alternative to Tweedie distribution variance structure to interpret collective risk models in actuarial applications. Shi (2016) proposes a copula-based multivariate Tweedie regression for modeling the semi-continuous claims while accommodating the association among different types, which also allows for dispersion modeling.

Another approach is the frequency-severity method, which needs to estimate the claim frequency and severity distributions separately and then multiply together relativities produced by each model. It is easier to communicate, which helps to greater understand business, and produces an option to include a variable in either frequency or severity. The compound loss distribution is a function of both loss frequency and severity. Quijano Xacur et al. (2011) compare the Tweedie approach against the frequency-severity approach and show that one important difference between these two methods is the variation of the scale parameter of the compound Poisson-gamma distribution when it is parameterized as an exponential dispersion model. Malavasi et al. (2022) propose a combination of regression models based on the class of Generalized Additive Models for Location Shape and Scale (GAMLSS) and a class of ordinal regressions. Denuit and Trufin (2017) bring up a collective approach to loss development by allowing more general severity distributions fitted to individual observations. Hua (2015) uses tail order of copulas to introduce tail negative dependence structure between loss frequency and loss severity, which improves the aggregate loss modeling. Frees et al. (2016) synthesize and extend the literature on multivariate frequency-severity regression modeling with a focus on copulas for modeling the dependence among outcomes. There is no preference between two approaches. If it is the first-time implementation, frequency and severity models are easier to find the pattern and signal of previous experience; if it is for model updates, pure premium is good to focus on until there has been a significant shift in your data. No matter which approach is used to deploy pipeline, the overall goal is to select a reasonable model that exacts signal out of historical experience that it is likely to be predictive of the future.

Our research focuses on the second approach: frequency and severity method. In recent years, more literature explore dependence structure of loss frequency and loss severity. Garrido et al. (2016) induce the dependence by treating the number of claims as a covariate in the model for the average claim size, so that pure premium is the product of a marginal mean frequency, a modified marginal mean severity, and a correction term. Lee and Shi (2019) propose a dependent modeling framework to jointly examine the two components in a longitudinal context where the quantity of interest is the predictive distribution. Jeong et al. (2021) compare the results by Garrido et al. (2016), Tweedie models, and the case of independence, and then demonstrate a superior model within GLMM framework.

Shi et al. (2015) propose conditional probability and copula approaches to correlate the number of claims and the average claim size in the conditional component. Chiou and Fu (2015) model crash frequency and severity, and accommodate spatial and temporal dependence by specifying a spatiotemporal function.

Most previous work focused on deriving closed form compound loss distribution by assuming that the frequency distribution has a closed from and the severity distribution is from the exponential family. The closed-form solutions are hardly obtained for cyber risk distributions; since the cyber risk has high frequencies and heavy-tailed severities, various pitfalls come from the existence of convolutions. Monte Carlo method has been raised to be successfully used for solving this problem. Ispirian et al. (1974) suggest a Monte Carlo method for calculation of the distribution of the ionization losses of charged particles passing through thin layers of matter. Septiany et al. (2020) provide the use of Monte Carlo method for selecting the distribution of claim frequency and claim severity. However, the following question is left unanswered: are there general and efficient algorithms that estimate compound loss distribution of non-parametric or dependent frequency and severity models? We give positive answers to this question by proposing MCMC algorithms to estimate compound loss distribution. In those algorithms, we first simulate the number of quarterly cyber incidents based on selected frequency model with posterior parameters, and then for each generated incident simulate the corresponding loss amounts (the number of data breached) based on selected severity model. The aggregation of all the loss amounts gives the compound loss distribution for that quarter. Our algorithms are designed to work for either parametric or semi-parametric distribution with closed or non-closed loss aggregation function.

1.4 Thesis Objectives and Outline

The purpose of the frequency analysis is to provide predictive analytics based on historical data on cyber incidents frequency aiming to help insurance companies examine, price and manage their cyber-related insurance risks. This analysis may be used by organizations as a reference in balancing their prevention costs with premiums according to their entity types and locations. We make use of related factors from cyber breach data and perform Bayesian regression techniques under Generalized Linear Mixed Model (GLMM). The key results of our loss frequency study are the following. Primarily, it is effective to use the complex NB-GLMM for analyzing the number of data breach incidents with uniquely identified risk factors such as type of breaches, type of organizations, and their locations. Additionally, it is practical to include in our model the notable correlation detected between the number of cyber incidents and the average severity amount (the number of data breached), as well as the time trend effects impacting the cyber incidents. Furthermore, it is efficient to use sophisticated estimation techniques for our analysis, including the Bayesian approach, MCMC method, Gibbs sampling, and Metropolis-Hastings algorithm. Ultimately, using the frequency-severity technique, it is feasible to use our predictive results for pricing the cyber insurance products with coverage modifications.

The objective of severity analysis is to generate a model that takes into account excess of zeros loss, spliced composites and mixture models under a global distribution with corresponding sets of covariates. Motivated by cyber risk specific nature, our study aims to fill these gaps using a finite mixture model (FMM) under a non-linear regression framework and a three-component splicing model with a zero-inflated component. Our zero-inflated mixture composite regression model (Zi-MCR) provides notable contributions overall to the actuarial literature as well as to the industry practice in developing/improving cyber insurance products. Our key findings reveal significant advancements in modeling cyber risks. Specifically, a flexible combination of mixture distribution model and splicing model is developed upon various candidate distributions, such as Gamma, Log-Normal, Weibull, Burr, Inverse Gaussian, and Pareto, effectively capturing the wide range and heavy-tailed nature of cyber loss severity. Additionally, we integrate FMM into a Generalized Linear Model (GLM) to utilize risk characteristics as covariates, allowing for the simultaneous estimation of GLM models for different subgroups and addressing individual risk characteristics. Moreover, we introduce a zero-inflated regression component to our model, enabling covariates to model the non-zero mixture distribution of the body and the extreme distribution of the tail, and the point mass zero rate, thus creating a comprehensive zero-inflated mixture and composite regression model with a complete cumulative distribution function. Finally, we provide a statistically rigorous method to quantify cyber risks under a single distribution that accounts for heavy tail nature of extreme losses, addressing the limitations of traditional insurance models that do not consider extreme values. This comprehensive approach enhances the accuracy and reliability of cyber risk modeling and prediction.

The rest of this thesis is structured as follows. Chapter 2 introduces the PRC chronology dataset to be studied in this thesis and presents the preliminary data analysis including descriptive statistics of dependent variable and regressors, exploratory analysis of utilized features and cluster analysis on geographical information. In Chapter 3, we propose a Bayesian negative binomial GLMM (NB-GLMM) for the quarterly cyber incidents recorded by PRC dataset. The quarter specific is one of the variations of random effects explained by the quarterly hierarchical panel data. Regression models on covariate predictors can capture variations of within-quarter heterogeneity effects. Moreover, GLMMs outperform the GLM by reveling features of the random effects distribution and allowing subject-specific predictions based on measured characteristics and observed values among different groups. Starting with introducing variable notations and distribution modeling structure in Section 3.1, we present the NB-GLMM for our breach data and the parameter inferences under Bayesian framework in Section 3.2. Section 3.3 shows the MCMC implementation and inference of the posterior distribution of parameters. The analysis of the PRC cyber breach chronology dataset using the NB-GLMM proposed is showed in Section 3.4. A simulation study and cross validation test against testing dataset to assess model performance are showed in Section 3.5. Finally, in Section 3.6 we discuss model applications and practical implications in cyber risk mitigation and management.

Chapter 4 presents a zero-inflated mixture and composite regression model (Zi-MCR) and discusses their application in cyber risk estimations. Section 4.1 reviews the definition of splicing models and finite mixture models, and propose our unique mixture and composite regression model adjusted by zero-inflated component based on dataset. Next, we introduce the expectation-maximization (EM) algorithm used to estimate coefficients and model parameters including E-step, M-step and specifications of some model parameters in Section 4.2. Followed by details on how to fit and choose from among these models as well as information about how to assess the goodness of fit of a model demonstrated by PRC data analysis in Section 4.3. Finally, we discuss applications of our model results from both the insurers' and potential insureds' perspective in Section 4.4. In Chapter 5, based on the results for our data-driven analysis presented in previous two chapters, we propose several approaches in generating aggregate losses and implementation strategies that can be utilized by the insurance industry. Starting with general notation and modeling structure of compound loss distribution in Section 5.1, we introduce a simple loss aggregation approach assuming that the loss frequency and severity are independent and the loss severity is not random in Section 5.2, and MCMC loss aggregation approach when the loss frequency is dependent on average loss severity and the loss severity has a specific zero-inflated mixture component distribution with parameters estimated based on the given data in Section 5.3. The impact of applying different deductibles, limits and reinsurance practice are discussed in Section 5.4. Finally, applications of the loss aggregation to current U.S. cyber insurance market are discussed in Section 5.5. Finally, we provide an overview of key contributions and innovations, highlighting the novel aspects introduced by this research, and also discuss the limitations encountered during the study, and outline potential directions for future research in Section 6.

Chapter 2

Description and Preliminary Analysis

In this chapter, we perform an empirical description and data analysis which support and motivate our data-driven modeling approaches and further analysis and application. Several necessary initialization procedures must be investigated. Starting with introducing frequently considered datasets in Section 2.1, we introduce PRC chronology and its statistical summary and patterns in Section 2.2. Followed by preliminary analysis in Section 2.3, where we investigate unique features of the dataset through an empirical data analysis and cluster analysis. Finally in Section 2.4, we discuss statistical challenges resulted from this dataset that our work is aiming to tackle, at the same time bring up possible improvements and updates for future work if more informative data is given.

2.1 Introduction

There are several publicly accessible cyber incidents datasets that are frequently considered and employed by researchers, such as the Open Security Foundation[1] DataLossDB (see, for example, Zeller and Scherer, 2021; Maillart and Sornette, 2010). Identity Theft Resource Centre[2] (ITRC) (Archer et al., 2012) and Verizon's Data Breach Investigations Reports[3] (DBIR), as well as Privacy Rights Clearinghouse (PRC) Data Breach Chronology (PRC, 2019). Our research is based on PRC data breach dataset which is to be introduced with details in the next section.

DataLossDB was founded in 2005 as an original data breach tracking project and operated until mid-2015, providing known and reported data loss incidents worldwide. This breach dataset includes the who, the when and the where, breach types, data type and data family. Driven by this dataset, Zeller and Scherer (2021) propose a new approach for modeling cyber risks using marked point processes and identify key covariates required to model frequency and severity of cyber claims. The resulting model is able to include the dynamic nature of cyber risk, while capturing accumulation risk in a realistic way. This paper also provides a comprehensive literature review on cyber risk and cyber insurance including data-driven studies, as well as data sources on data breaches. In an earlier study, Maillart and Sornette (2010) investigate some noticeable statistical properties of cyber-risks based on DataLossDB dataset, which are used to quantify the distribution and time evolution of information risks on the Internet. Their findings help understand the underlying mechanisms and thus present opportunities for risks mitigate, control, predict and insure them at a global scale.

ITRC provides superior support to victims at no charge to consumers in the U.S., and educate consumers, business entities and organizations on best practices for fraud and identity theft detection, reduction and mitigation. This site keeps data breaches information including company name, state in the U.S., breach category and number of records exposed when the incident occurs. Based on this dataset, Archer et al. (2012) introduce a general model describing the identity theft and fraud process including an explanation of various components that make up this process model and potential crimes resulting from the criminal activities.

VERIS Community Database (VCDB) represents a broad ranging public effort to gather cyber security incident reports in a common format. The collection is maintained by the Verizon RISK Team, and is used by Verizon in its highly publicized annual Data Breach

[1] The Open Security Foundation DataLossDB, https://www.datalossdbdotorg.wordpress.com/

[2] Identity Theft Resource Centre Data Breaches, https://www.idtheftcenter.org/category/blog/data-breaches/

[3] 2023 Data Breach Investigations Report, https://www.verizon.com/business/resources/reports/dbir/

Investigations Reports (DBIR). Seh et al. (2020) conduct an in-depth analysis of healthcare data breaches based on DBIR and draw inferences from it, and thereby use the findings to improve healthcare data confidentiality. Liu et al. (2015) characterize the extent of cyber security incidents referenced by Verizon DBIR and make predictions based on externally observable properties of an organization's network.

PRC database records cyber breach incidents between years 2001 and 2022. Most of the breach data comes from state attorneys general and the U.S. Department of Health and Human Services. This dataset contains the data breach incidents as well as the number of records breached due to these breach incidents. The dataset serves as a resource for researchers and practitioners examining the effect of data breaches on the performance of insurance companies. Our study is based on the latest available PRC data breach chronology downloaded with 9012 breach observations happened in the United States since year 2001. After removed incomplete and inconsistent observations, 8095 incidents including 4161 medical incidents and 3934 non-medical incidents are investigated and analyzed. We restrict the sample to the time period from 2001 since cyber risk only becomes a serious issue in the 21st century and the data in the last century are very sparse.

2.2 Data Breach Chronology Database

We have presented in the last section several frequently considered databases from nonprofit corporations and some studies based on them. Our research is primarily driven and based on Privacy Rights Clearinghouse (PRC) Data Breach Chronology database. In this section, we perform an empirical data analysis which supports and motivates our data-driven modeling approaches and further analysis and applications. Several necessary initialization procedures are investigated. Starting with the explanatory data analysis, we investigate unique features of this dataset through an empirical data analysis, followed by a cluster analysis to study the multidimensional location feature of this dataset.

PRC is a nonprofit organization aiming to provide the most accurate and up-to-date information, which stimulates research in cyber related loss modeling and prediction, as well as developing associated insurance products and their premium determination. The PRC dataset has widely been studied by several research works from various perspectives. For example, Edwards et al. (2016) develop Bayesian generalized linear models to investigate trends in data breaches. Their analysis shows that none of the size and the frequency of data breaches has increased over the past decade, and both are heavy-tailed. Furthermore, they find that the daily frequency of breaches can be modeled by a negative binomial distribution, while the size of data breaches can be described by the log-normal family of distributions. Eling and Loperfido (2017) investigate this dataset under statistical and actuarial science framework by using multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breach information. They show that different types of data breaches need to be modeled as distinct risk categories and provide useful insights for actuaries working on the implementation of cyber insurance policies.

The data is recorded under case unit with breach types, business types, incident entities and their geological location; these variables could be valuable predictors while generating regression models and making predictions. Table 2.1 shows a sample of data breach incidents happened in 2018, where the type of cyber event and the victim's information, such as the company's name, type of business, and location, are gathered from breach incidents; see Table 2.2 for abbreviations used in this table. It is worth mentioning that a breach incident happened to Epsilon corporation in Texas, Jan 2011 caused the largest number of records which is 250 millions. Because it contains risk-related characteristics that can be utilized as rating factors, this information is essential for filing insurance rates, and therefore is fully utilized in our studies for both the frequency and severity of the data breaches. In Chapter 3, a generalized linear mixed model (GLMM) is proposed to study the quarterly frequency (number of incidents) of the data breaches recorded in this PRC database and its application to the cyber insurance is discussed. In Chapter 4, we are interested in the number of records breached by each recorded data breach incident collected in PRC database, which is considered as the severity of the breach caused by cyber breach incidents. We late convert the breached data record to dollar amount loss in order to get a dollar amount magnitude.

Incident Date	Type of Breach	Type of Business	Location	Loss of Records
2018/02/03	CARD	BSF	California	30
2018/05/26	HACK	GOV	Washington	1000
2018/06/30	DISC	MED	Massachusetts	900
2018/09/27	PHYS	EDU	Florida	1500
2018/10/09	INSD	BSR	Texas	700
2018/12/05	PORT	NGO	Ohio	150

Table 2.1: Sample of PRC Chronology

In Chapter 3, we model PRC quarterly counts (the number of data breach incidents) as a function of breach type, breach entity and location, which can be linear predictors of target variable via general design matrices. Moreover, we model relationships among risk exposure characteristics through matrix design by taking all featured combinations as different risk exposures. In order to lower the dimension of parameter matrix, reduce the volatility of data and stable the rates overtime, we further combine levels with similar information into new representative levels of three categorical variables under clustering analysis (Jain et al., 1999): South, West, Northeast and Midwest (according to U.S. Census Bureau) under location, external and internal under breach type, and business and non-business under organization type for non-medical organizations as showed in Table 2.2. Note that unknown types of breach and business have been eliminated due to their incomplete information.

Original Types		Combined Levels
MED	Healthcare, Medical Providers and Insurance Services	Medical
BSF	Businesses (Financial and Insurance Services)	Business
BSO	Businesses (Other)	
BSR	Businesses (Merchant including Online Retail)	
EDU	Educational Institutions	Non-business
GOV	Government or Military	
NGO	Nonprofits	
CARD	Fraud Involving Debit and Credit Cards	External Malicious
HACK	Hacked by an Outside Party or Infected by Malware	
INSD	Insider	Internal Malicious
PHYS	Physical	
PORT	Portable Device	
STAT	Stationary Computer Loss	Internal Negligent
DISC	Unintended Disclosure	

Table 2.2: Covariate level combinations

As a result, the original case unit basis dataset is manipulated as a hierarchical dataset with quarterly counts on uniquely identified 16 level combinations. These combinations divided the dataset into three dimensional augmentations. Besides targeting counts variable and designing covariate matrix described above, it is worth mentioning the following features of the PRC empirical breach frequency distribution. Figure 2.1 shows the empirical quarterly counts between years 2001 and 2018 density performance of non-medical organizations (left) and medical organizations (right). Frequency counts are aggregated on quarterly interval of specific combination subjects. Both plots reflect the fact that there exists a portion of zero incidents and the data is dispersed over a wide range. It is noteworthy that, although density plots for non-medical and medical organizations share overall similarities, the detailed performances between two plots are different showing the cyber related risk nature differences between the non-medical and medical organizations. For instance, the proportion of zeros is much higher for non-medical organizations and the scale for non-medical

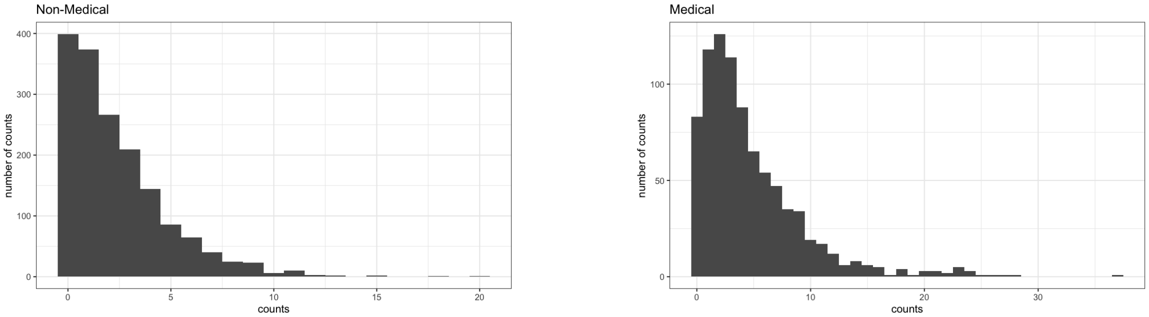


Figure 2.1: Histograms on different organizations

empirical distributions is more centered. These observations follow the current trending that medical identity theft and medical data breaches are vividly rising at disproportionate rates compared with other attacked industries (Rathee, 2020). The NAIC (2020) Cybersecurity Report points out that healthcare breaches grew by 33.3% higher than the data breach growth rates from other type of organizations. In addition, frequency distribution of medical organization shows heavier tails by having more large loss amount breach incidents. All these suggest that it may be necessary to separately analyze of data breaches happened to the non-medical organizations and that to the medical organizations.

2.3 Preliminary Analysis of PRC Chronology Data

2.3.1 Descriptive Analysis of Data Breach Chronology

In this subsection we perform exploratory data analysis on breach incident counts (frequency) that helps gain insights into the distribution of our target variable. Table 2.3 displays summary statistics of the quarterly number of breach incidents that the non-medical and medical organizations incurred between years 2001 and 2018. The incidents of the medical subset is more widespread ranging from 0 to 37, whereas that of the non-medical ranges from 0 to 20 only. Both of them are right skewed with mean greater than median and the medical subset has a heavier tail and shows overdispersion with a large variance. Both quarterly count frequencies contain a proportion of zeros which means some characteristic combinations do not incur breach incidents at these quarters.

Entity Type	Minimum	Maximum	Mean	Median	Variance	Proportion of Zeros
Non-Medical	0	20	2.277	2	6.014	0.267
Medical	0	37	4.762	3	22.274	0.096

Table 2.3: Summary statistics of quarterly frequency counts

With these features, we fit the Poisson, negative binomial (NB), zero-inflated Poisson and zero-inflated NB distributions to the medical and non-medical counts subdatasets, respectively. While the Poisson and NB distributions are commonly used in modeling the claim counts in actuarial field, the NB distribution could be a conservative model choice as it can handle overdispersion and its zero-inflated version could be appropriate due to the appearance of heavy zeros observed in the non-medical subdataset. When several models are available, one can compare the model performance based on statistical likelihood measures; here we use AIC (Akaike information criterion Bozdogan (1987)). AIC is one of the most popularly used measures, for these above-mentioned distributions, in order to testify which distribution preliminarily describes best the breach incident frequencies. It penalizes a model with larger number of parameters and is defined as

$$\text{AIC} = -2\,\text{LogLik} + 2p$$

where LogLik denotes the logarithmic maximum value of likelihood function of fitted model and p is the number of model parameters. A relatively small value of AIC is favorable for the fitted model.

Table 2.4 shows the AIC values for the distributional models that we fit. Based on these values, we find that the NB model fits both the medical and non-medical data best. Our findings actually coincide with the conclusions from several studies of cyber incidents in literature. For example, Edwards et al. (2016) model the frequency of data breaches with the NB distribution under Bayesian approach. Joe and Zhu (2005) provide helpful insights, besides the likelihood metrics, in selecting a better fitting NB distribution for modeling count data with long right tails. Proceeding along similar lines, we adopt the NB as the target regression distribution of GLMM model based on natures of PRC dataset, which is discussed in Chapter 3.

Entity Type	Non-Medical	Medical
Poisson	7617	5947
Negative Binomial	**6739**	**4552**
Zero-inflated Poisson	7165	5657
Zero-inflated Negative Binomial	6941	4555

Table 2.4: AIC values for tested distributions fitting quarterly loss counts

2.3.2 Exploratory Analysis of Data Breach Patterns

As discussed previously in Chapter 1, fitting an adequate loss distribution to the cyber breach dataset is difficult due to its nature. Here, we conduct an empirical data analysis of related target and explanatory variables on the PRC data set to demonstrate the necessity of addressing/accounting for several risk features. A summary statistics of this data set is provided in Table 2.5.

	Number	Zero Prop.	$q_{0.25}$	Mean	Median	$q_{0.75}$	Maximum
Total	8095	32.90%	1000	1018500	2800	13000	5×10^8
Medical	4161	15.66%	1000	69400	2300	8800	7.88×10^7
Non-medical	3934	51.12%	900	2750400	4700	38900	5×10^8

Table 2.5: Summary statistics of PRC loss records

The first row of Table 2.5 provides summary statistics for the target variable, the "number of records" breached (loss severity) from a total number of 8095 data breach incidents, rounded to the nearest 100 units, where q_α denotes the empirical α-quantile of positive losses. We observe from these summary statistics that the number of records has a 32.9% excess of zeros and a very heavy right tail, given that the sample mean is significantly larger than the sample median. The heavy tail nature of the data can be revealed by the fact that some types of loss such as competitive advantage and reputation damage, can occur infrequently but with extreme severity, leading to disproportionately large impacts on the overall risk profile. The breach incident is recorded as no reported loss or expenses, if there were no records lost in that incident or damage can not be measured financially. The breached records range from 0 to 500 million which is difficult to model using one distribution. In this regard, our analysis of the loss amounts is based on the logarithm of severity in order to maintain complete low and high loss amounts information.

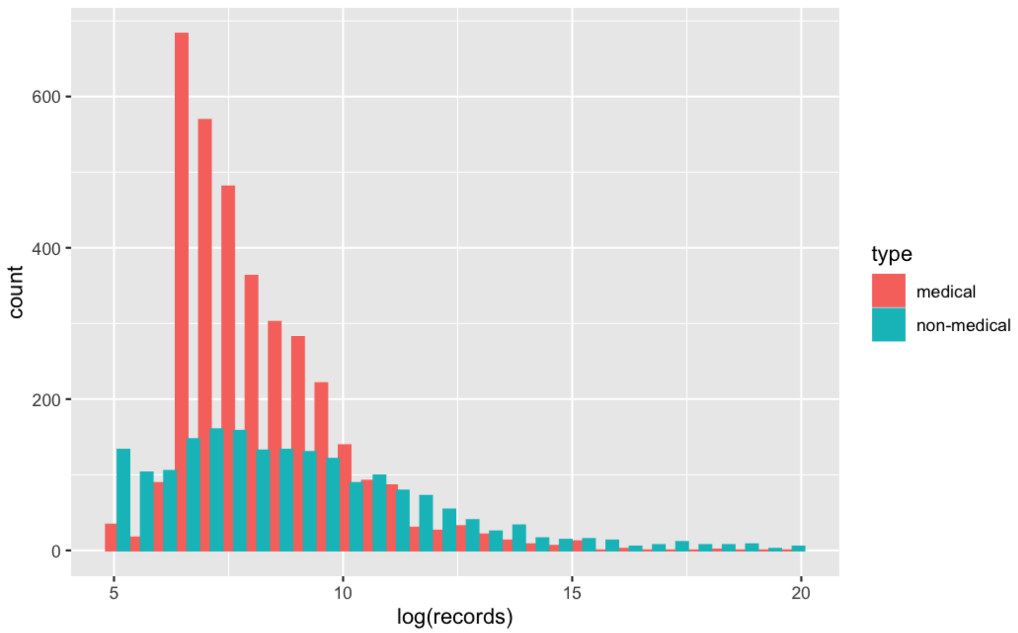


Figure 2.2: Histograms between medical and non-medical organizations

The PRC data set contains three explanatory variables that can be used as regressors: breach type, organization type, and company location. The first two variables are documented to have seven subcategories each, while the location is listed in 50 geographical states. We modify on their levels, based on their nature and characteristics to reduce factor dimensions and increase predictive power. Table 2.2 summarizes the combined model inputs of business and breach types. The level combination of the geographic locations is discussed and described in Section 2.3.3. After obtaining six combined levels of information regarding medical, business, non-business, external malicious, internal malicious and internal negligent, we investigate their performance on the target variable and find that those medical and non-medical organizations behave differently concerning the number of breached records. It can be observed from last two rows of Table 2.5 the significant differences between the medical data and the non-medical data, which covers business and non-business types of organizations as in Table 2.2. Although they all show the similar pattern that processes point mass of zero, two zero proportions differ notably to the extent of the heavy tail and maximum amount measured on the non-zero loss amount. In addition, the medical losses are more compact compared to non-medical losses. We hence postulate that the underlying severity distribution features multi-modality; in this sense, a multi-modal distribution or mixture distribution could be candidates for modeling the overall losses.

The above-mentioned fact can also be observed from Figure 2.2, where both histograms of logarithmic records breached and incurred by the medical and the non-medical organizations are displayed. The non-zero severity body part of empirical density of medical organizations has a peak of around 600 records, and the probability for losses being smaller than the mode value is relatively low. Meanwhile, the body part density for non-medical organizations shows a relatively smooth and flat distributional pattern before and after its mode point and relatively a heavier tail. Compared to financial services industry, which has spent the last 20 years focusing on cyber security and protection (Bell and Ebert, 2015), healthcare organizations are not as frequently attacked by cyber related incidents. Medical organizations form traditionally risk retention groups to mitigate huge liability losses caused by cyber breaches, making them reluctant to understand, track, report, and manage threats via open market cyber insurance coverage. Besides, mature incident and vulnerability risk management processes are lacking in most medical organizations (Williams and Woodward, 2015). Thus, daily threats are not even reported or managed effectively, which explains the low occurrence of cyber-severity losses of less than 600 breached records. Even though some of the distributions are already appropriate to model losses with heavy tails, they do not account for this multi-modality case resulting from data variations observed between medical and non-medical organizations. In this regard, estimating the moderate loss density component with a fixed number of mixed components is advantageous.

2.3.3 Cluster analysis

Because the PRC data also contains the geographical location of the victims of cyber attacks, a list of 51 states of U.S. with their latitudes and longitudes serves as the raw data information. It is a common practice that the number of levels in the geographical rating factor are to be reduced in order to provide effective risk measurement for insurance rate-making. For this purpose, we use one of the initialization strategies, cluster analysis (Roberts, 1997), to do the analysis. Clustering analysis is a newly developed computer-oriented data analysis which utilizes unsupervised machine learning algorithms to segment a data set based on similarities between the data points. Hofstetter et al. (2014) clarify the use of cluster analysis and factor analysis, and provide a guideline to a universal understanding of the analysis of co-occurrence of risk behaviors. Zheng et al. (2014) apply K-means cluster analysis to classify the near-crash cases into different driving risk levels in the vehicle kinetic energy.

We conduct cluster analysis for three reasons. First, it avoids diluting the predictive power caused by the geographical location factor with 51 levels. Second, when states with similar characteristics are grouped, implementing rate-making is simpler. Third, it reduces the likelihood that the rate for one area is drastically differ from that for its neighbouring areas. Cluster analysis divides observations into distinct groups so that the observations within each group are quite similar to one another, as opposed to grouping 51 states into some official government regions, such as those used by the U.S. Census Bureau and the Standard Federal Regions. Before clustering, we conduct a cluster analysis using the means of latitude and longitude in each state as representatives. In this regard, we smooth the regression coefficients to make them more reasonable and interpretable, given that clustered groups are based on state average level. Now we have a set of 8095 observations, each with two features, longitude and latitude, that can be used to identify subgroups. We attempt to discover geographical heterogeneity structures based on the PRC data set, which is an unsupervised problem.

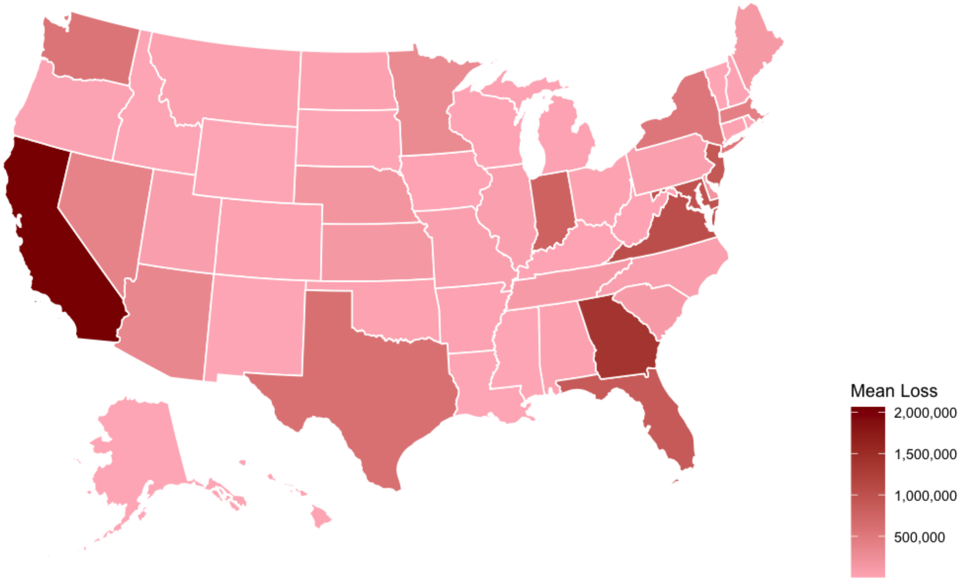
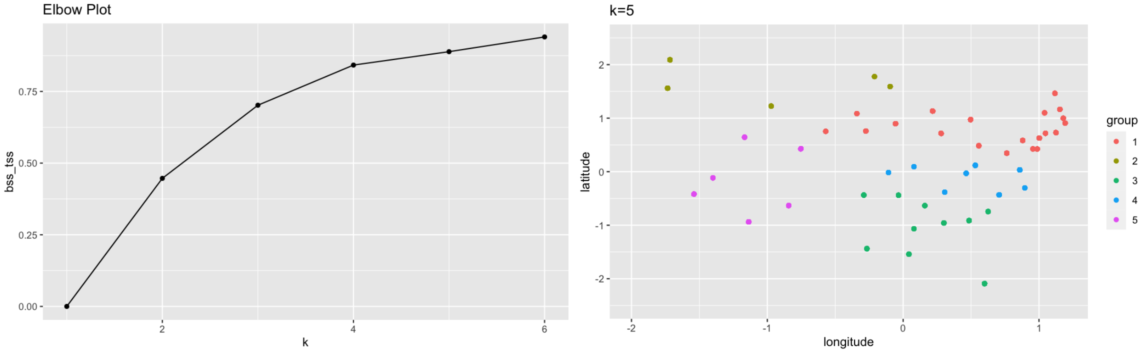


Figure 2.3: Average severity level among states

Figure 2.3 represents the geographical heat map information in a two-dimensional space of longitude and latitude. These are the first two principal components of the data, which summarize the location information of in total 8095 investigated incidents in terms of two geographical dimensions. Each small and closed area corresponds to one of the 51 states, allowing for a visual examination of the average severity level for signs of clustering. There appear to be multiple groups of clusters with similar colour patterns. Two commonly used clustering techniques are K-means (Likas et al., 2003) and hierarchical (Johnson, 1967), which have been widely applied in territory studies for finding patterns and investigating the underlying geographical structure of the data. This study uses the K-means method with elbow (Bholowalia and Kumar, 2014) to show the K-means performance and to find an efficient and effective K. The elbow method is a default standard method for determining the



(a) Elbow plot for clusters

(b) Five geographical clusters

Figure 2.4: Cluster selection

optimal number of clusters for a characteristic process. The K-means clustering algorithm formalizes finding the best similarity grouping where the variations among observations within each cluster are as small as possible, and the variation between clusters is significant. The similarity is measured by the error sum of squares (SSE) (also called squared Euclidean distance) (Agrawal et al., 1993), one of the most widely used cluster distance criteria:

$$\mathrm{SSE}_{ij} = \sum_{k=1}^{K} (x_{ki} - \bar{x}_{k\cdot})^2 + \sum_{k=1}^{K} (x_{kj} - \bar{x}_{k\cdot})^2,$$

where i and j are two dimensions representing variable combinations and number of quarters, and $x_{k\cdot}$ represents the kth cluster of the K clusters whereas $\bar{x}_{k\cdot}$ represents the mean distance of group k. We manually conduct a K-means cluster analysis with one to six clusters and calculate the ratio between individual cluster sum of squares and the total sum of squares for each round. We take this ratio as the y-axis and create an elbow plot as illustrated in Figure 2.4(a). The plot demonstrates the elbow at $K = 5$, beyond which the gains in between cluster's sums of squares appear to be minimal because the increase in total sum of squares after $K = 5$ is greatly shrinking down; therefore five is the best cluster cut-off point. Figure 2.4(b) depicts the relative geographical location of five clusters, while Appendix A provides context-specific information about cluster partitioning by state. By this way, we can identify the geographical segments of cyber severity and classify them according to similar risk factors.

2.4 Statistical Challenges

PRC chronology focuses on events that occur to legal entities instead of individuals, and contain two major types of breaches. The first type of incidents is resulted from external activities such as card fraud and hacking. The second type of incidents is related to internal operational activities such as insider employees, physical documents and portable device, and unintended disclosure. By studying this dataset, we aim to find the general pattern of cyber risks that is persistent across different sources and categories.

Most significant impacts to the PRC chronology data quality come from Amounts of Dark Data, Number of Empty Values, and Data Time-to-Value and metrics. Even though the chronology contains different kinds of fields with information related to breach incidents, such as address, reporting source and website, the useful fields that contain information which is not highly correlated or homogeneous are relatively scarce. Excluding the dark fields that can not be used as signal variable in the model, we narrow down to five meaningful features in constructing our models, date of breach, number of affected digital records, type of breaches, type of organizations and geographical location. Since we utilize those features in generating regression models, the robustness of our models relies heavily on proportion of non-empty values. After cleaning the data and using the sample after year 2001, we have only 8095 incidents with known number of record breached, types related features and location.

Furthermore, we started to explore this dataset in 2019 and have been working on it since then. Although PRC continued to update their chronology database until 2021, we decide not to include additional two years sample due to the data quality concerns. As we have seen, the global COVID-19 pandemic (that was declared by the World Health Organization on March 11, 2020) has the significant impact to all the industries as we mentioned previously. Hence, different organizations may face different cyber risks as before, and incidents happened during the pandemic period may not be observed consistently with their pre-pandemic patterns.

Last but not least, it is worth mentioning that another limitation coming from the chronology in actuarial science perspective is that there is no exposure information that can be measured or indirectly derived from this dataset. This dataset contains more than 10000 breach events that are reported by victims or census institutions, and records are documented in an incurred and reported basis instead of tracking an amount of entitles' breach activities within a time period. Thus, there is no classification for risk exposure and the number of digital assets exposed within a given time period, that we can extract directly from the data. This results in treating 8095 observations with full exposure within quarterly modeled period. Therefore, the analysis of risk frequency can be based on quarterly number of incidents, and risk severity is investigated on individual observation basis.

Chapter 3

Generalized Linear Mixed Model for Cyber Loss Frequency Analysis

In this chapter, we propose a Bayesian negative binomial generalized linear mixed model (NB-GLMM) for the quarterly cyber incidents recorded by PRC dataset. The quarter specific is one of the variations of random effects explained by the quarterly hierarchical panel data. Regression models on covariate predictors can capture variations of within-quarter heterogeneity effects. Moreover, GLMMs outperform the generalized linear model (GLM) by reveling features of the random effects distribution and allowing subject-specific predictions based on measured characteristics and observed values among different groups. Starting with introducing variable notations and distribution modeling structure in Section 3.1. We present the NB-GLMM for our breach data and the parameter inferences under Bayesian framework in Section 3.2. Section 3.3 shows the Markov chain Monte Carlo (MCMC) implementation and inference of the posterior distribution of parameters, followed by analysis of cyber breach chronology dataset as modeling illustration and application in Section 3.4. In order to evaluate model robustness, a simulation study and cross validation test against testing dataset to assess model performance are showed in Section 3.5. Finally, we discuss model applications and practical implications in cyber risk mitigation and management in Section 3.6; detailed discussion of aggregated total claim costs and cyber insurance applications are presented in Chapter 5. The research presented in this chapter, including the methodology, has been published in Sun and Lu (2022).

3.1 Notation and Model Formulation

In this section, we first introduce notations before a GLMM (McCulloch, 2006) for modeling the quarterly number of data breaches is formulated for our study. Assume that the total number of risk combinations is I and the total number of quarters is J. Let N_{ij} be a random variable representing the number of data breach incidents of ith combination in jth quarter, where $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$. Let μ_{ij} be the mean of N_{ij} conditional on $\boldsymbol{\beta}_j$ and $\boldsymbol{b}$, where $\boldsymbol{\beta}_j = (\beta_{1,j}, \beta_{2,j}, ..., \beta_{H,j})^T$ is a H-dimensional vector of regression coefficients for the jth quarter, and $\boldsymbol{b} = (b_1, b_2, ..., b_G)^T$ is a G-dimensional vector of regression coefficients. Furthermore, let $\boldsymbol{x}_{ij} = (x_{1,ij}, x_{2,ij}, ..., x_{H,ij})^T$ be a H-dimensional vector and $\boldsymbol{z}_{ij} = (z_{1,ij}, z_{2,ij}, ..., z_{G,ij})^T$ be a G-dimensional vector, which are measured covariates for the ith combination in the jth quarter.

Assume that $\{N_{ij}, i = 1, 2, \dots, I\}$ are conditionally independent for fixed j with given $\boldsymbol{\beta}_j$ and $\boldsymbol{b}$, and follow a distribution with probability density function $f(\cdot|\boldsymbol{\beta}_j, \boldsymbol{b})$ and mean $\mu_{ij}, i = 1, 2, \dots, I$, respectively. Let $g(\cdot)$ be a link function. Then our model can be described as follows:

$$\begin{aligned} N_{ij}|\boldsymbol{\beta}_j, \boldsymbol{b} &\sim f(n_{ij}|\boldsymbol{\beta}_j, \boldsymbol{b}), \qquad i = 1, 2, \dots, I.\ j = 1, 2, \dots, J \\ \mathrm{E}[N_{ij}|\boldsymbol{\beta}_j, \boldsymbol{b}] &= \mu_{ij}, \\ g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T\boldsymbol{b}, \\ \boldsymbol{\beta}_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \qquad j = 1, 2, \dots, J \end{aligned} \tag{3.1}$$

in which the heterogeneity among the regression coefficients $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J$ is described by a multivariate normal distribution with mean $\boldsymbol{\theta}$ and a variance-covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ii} = \sigma_i^2$. Note that random vector variable $\boldsymbol{\beta}_j$ reflects the within group variations for the jth group (quarter), while the i.i.d. multivariate normal random vector variables $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J$ reflect the between group variations for total of J groups (quarters).

In fact, the model (3.1) can be written as a standard GLMM format (McCulloch, 2006). Let $\boldsymbol{\eta}_j = (\eta_{1j}, \dots, \eta_{Ij})^T$, and

$$\boldsymbol{X}_j = \begin{bmatrix} x_{1,1j} & x_{2,1j} & \cdots & x_{H,1j} \\ x_{1,2j} & x_{2,2j} & \cdots & x_{H,2j} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1,Ij} & x_{2,Ij} & \cdots & x_{H,Ij} \end{bmatrix}, \qquad \boldsymbol{Z}_j = \begin{bmatrix} z_{1,1j} & z_{2,1j} & \cdots & z_{G,1j} \\ z_{1,2j} & z_{2,2j} & \cdots & z_{G,2j} \\ \cdots & \cdots & \cdots & \cdots \\ z_{1,Ij} & z_{2,Ij} & \cdots & z_{G,Ij} \end{bmatrix}.$$

Write $\boldsymbol{\beta}_j = \boldsymbol{\theta} + \boldsymbol{u}_j$, where $\boldsymbol{u}_j$ is a H-dimensional vector. Then the explanatory variable structure $\boldsymbol{\eta}_j$ given in (3.1) can be rewritten as a sum of fixed effects and random effects components via the treatment design (Stroup, 2012):

$$\begin{aligned} \boldsymbol{\eta}_j &= \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{Z}_j\boldsymbol{b} \\ &= \boldsymbol{M}_j\boldsymbol{\gamma} + \boldsymbol{X}_j\boldsymbol{u}_j, \\ \boldsymbol{u}_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{3.2}$$

where $\boldsymbol{M}_j = [\boldsymbol{X}_j, \boldsymbol{Z}_j]$ is a $I \times (H+G)$ covariate matrix and $\boldsymbol{\gamma} = [\boldsymbol{\theta}^T, \boldsymbol{b}^T]^T$ is a $(H+G)$-dimensional vector. Clearly, in (3.2) $\boldsymbol{M}_j\boldsymbol{\gamma}$ represents the fixed effects component of the mean vector, while $\boldsymbol{X}_j\boldsymbol{u}_j$ represents the random effects component of the mean vector, for which a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ is assigned to $\boldsymbol{u}_j$ for all j. This shows that between group effects and within group effects can be separated for a given information about the hierarchical data.

As suggested by our empirical study showed in Section 2.3.1, we assume that N_{ij} given $\boldsymbol{\beta}_j$ and $\boldsymbol{b}$ follows a NB distribution with mean μ_{ij} and dispersion parameter ξ_j, and a log link is used, namely, for $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$ and $n_{ij} = 0, 1, \ldots$

$$f(n_{ij}|\mu_{ij}, \xi_j) = \frac{\Gamma(n_{ij} + \xi_j^{-1})}{\Gamma(\xi_j^{-1})\Gamma(n_{ij}+1)} \left(\frac{1}{1+\mu_{ij}\xi_j}\right)^{\xi_j^{-1}} \left(\frac{\mu_{ij}}{\xi_j^{-1} + \mu_{ij}}\right)^{n_{ij}}, \tag{3.3}$$

where μ_{ij} is the mean of N_{ij} as denoted in (3.1) such that $\ln(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T\boldsymbol{b}$, and $\xi_j(>0)$ is also called the shape parameter used in the variance expression of N_{ij}, which is $\mu_{ij} + \xi_j\mu_{ij}^2$. Here we take the type II NB distribution, termed as NB2 (Hilbe, 2011) due to the quadratic natural of its variance function. The NB2 distribution can be generated from the Poisson-gamma mixture model and is also a member of exponential family. This formulation is adopted because it allows the modeling of within group heterogeneity using a gamma distribution.

In our data breach frequency data analysis, the recorded information from the PRC dataset on the type of breaches, type of organizations and entity location, when a data breach incident occurs, are used as covariates. Based on our further exploration and evidence observed on this dataset showed in Section 3.4, we also take into consideration the variations in average severity (the number of data breaches caused by data breach events) of each combination and the time trend. We consider the parameters corresponding to type of breaches, type of organizations, entity location and average severity as both fixed and random effects, and consider the parameters for time trend as fixed effects. We thus have $H = 6$ for $\boldsymbol{x}_{ij}$ and $\boldsymbol{\beta}_j$, and $G = 3$ for $\boldsymbol{z}_{ij}$ and $\boldsymbol{b}$ under cubic polynomial assumption for the time trend; the corresponding dimension of fixed effects covariates (type of breaches, type of organization, location, average severity, time trend) in (3.2) is thus nine and that of random effects covariates (type of breaches, type of organization, location, average severity) is six. Instead of letting only one covariate contains random effects, we consider that the random effects rely on all the risk characteristic features derived from raw factors. Besides hierarchical structure variations, the time trend effects are considered as fixed effects in the portion of the mean of GLMM. We then investigate unknown parameters under Bayesian framework combined with prior and posterior distributions. Finally, we introduce parameter inferences on hyper parameters using Markov chain Monte Carlo (MCMC) and Metropolis-Hasting (M-H) algorithms. More details on the GLMM for the PRC frequency dataset are presented in Section 3.4.

3.2 Estimation Procedures under Bayesian Framework

The GLMM has been specified in Section 3.1. We now in this section consider the inferences about the built-in process that generates the data. There are various ways to approximate the likelihood used for estimating GLMM parameters, including pseudo and penalized quasilikelihood (PQL) (see, for example, among others, Schall, 1991; Wolfinger and O'connell, 1993; Breslow and Clayton, 1993), Laplace approximations (Raudenbush et al., 2000), Gauss-Hermite quadrature (GHQ) (Pinheiro and Chao, 2006) and MCMC algorithms (Gilks, 1996). First three methods explicitly integrate over random effects to compute the likelihood, whereas the MCMC method generates random samples from the distributions of parameters for fixed and random effects. We adopt the MCMC method in this study, because it can be easily used in considering multiple random effects on part of explanatory variables for our dataset. MCMC algorithms are normally used under a Bayesian framework which incorporates prior information based on previous knowledge about the parameters or specifies uninformative prior distributions to indicate the lack of knowledge. Parameter estimations are made through the posterior distribution that is computed using Bayes' theorem, which is the cornerstone of Bayesian statistics and provides an effective approach in making inferences (Dempster, 1968).

3.2.1 Prior and posterior distribution

In addition to Bayesian flavor and well posed statistical model, MCMC involves possibly challenging technical details including choosing appropriate priors and efficient algorithms for granular problems. The Bayesian approach also requires the specification of prior distributions of all model parameters. Note that in Bayesian GLMM analysis, it normally assumes that the prior distribution of coefficient vector is multivariate normal distributed and the variance-covariance matrix is inverse Wishart distributed. Under our model described by (3.1), the prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are assumed and their posterior distributions are discussed in the following.

We first present the prior and posterior distribution of $\boldsymbol{\theta}$ assuming that the variance-covariance matrix $\boldsymbol{\Sigma}$ is known. Suppose that the mean vector $\boldsymbol{\theta}$ is multivariate normal distributed with mean vector $\boldsymbol{\mu}_0$ and variance-covariance matrix $\boldsymbol{\Lambda_0}$, that is,

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda_0}),$$

which is actually a conjugate prior distribution of $\boldsymbol{\theta}$, and it is well known that the corresponding posterior distribution is also multivariate normal distributed. Following Hoff (2009), the full conditional (posterior) distribution of $\boldsymbol{\theta}$, given a sample of regression coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$ and $\boldsymbol{\Sigma}$, can be easily derived as

$$[\boldsymbol{\theta} \,|\, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma}] \sim \mathcal{N}(\boldsymbol{\mu}_J, \Lambda_J), \tag{3.4}$$

where $\boldsymbol{\mu}_J$ is the conditional mean vector and Λ_J is the variance-covariance matrix, given by

$$\begin{aligned}\boldsymbol{\mu}_J &= (\Lambda_0^{-1} + J\boldsymbol{\Sigma}^{-1})^{-1}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + J\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\beta}}),\\ \Lambda_J &= (\Lambda_0^{-1} + J\boldsymbol{\Sigma}^{-1})^{-1},\end{aligned}$$

in which $\bar{\boldsymbol{\beta}} = \left((1/J)\sum_{j=1}^{J}\boldsymbol{\beta}_{1j}, \ldots, (1/J)\sum_{j=1}^{J}\boldsymbol{\beta}_{Hj}\right)^T$ is a H-dimensional vector average.

We now discuss the prior and posterior distribution of $\boldsymbol{\Sigma}$. Having information of $\boldsymbol{\Sigma}$ helps in detecting group variance caused by group specific features, especially the relationship between covariates which could be evaluated with correlation coefficient $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_i^2\sigma_j^2}$. In Bayesian statistics, in the context of the multivariate normal distribution, the Wishart distribution is the semi-conjugate prior to the precision matrix $\boldsymbol{\Sigma}^{-1}$ (Chatfield and Collins, 2018), and hence the inverse-Wishart distribution is the semi-conjugate prior distribution for the variance-covariance matrix $\boldsymbol{\Sigma}$. Assume now a conjugate inverse-Wishart prior distribution for $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}\left(\nu_0, \boldsymbol{S}_0^{-1}\right),$$

where ν_0 is a scalar hyper-parameter and $\boldsymbol{S}_0^{-1}$ is a symmetric $H \times H$ positive definite matrix. Based on (3.1) that regression coefficients $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$, are multivariate normal distributed, the conditional posterior distribution of $\boldsymbol{\Sigma}$, given a sample of regression coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$ and $\boldsymbol{\theta}$, can be written as

$$[\boldsymbol{\Sigma} \,|\, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\theta}] \sim \mathcal{W}^{-1}\left(\nu_0 + J, [\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}}]^{-1}\right) \tag{3.5}$$

where $\nu_0 + J$ is the hyper-parameter and $[\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}}]^{-1}$ is the covariance matrix, in which $\boldsymbol{S}_{\boldsymbol{\theta}}$ is the matrix of residual sum of squares with respect to mean vector $\boldsymbol{\theta}$, given by

$$\boldsymbol{S}_{\boldsymbol{\theta}} = \sum_{j=1}^{J}(\boldsymbol{\beta}_j - \boldsymbol{\theta})(\boldsymbol{\beta}_j - \boldsymbol{\theta})^T.$$

Detailed derivations can be found in Hoff (2009).

3.2.2 Maximum likelihood estimation of dispersion parameter

The maximum likelihood estimation for the dispersion or heterogeneity parameter from a NB distribution is discussed with details in Piegorsch (1990). Under our GLMM setting, ξ_j is the dispersion parameter for jth quarter in (3.3) which scales the population variance. In our model, the generalized linear regression algorithm on target NB2 distribution with a log link function leaving heterogeneity parameter to be entered into GLMM model as a constant (Hilbe, 2011). As it can be seen in the estimation algorithm presented in the next section, parameter $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_J\}$ are estimated outside and subsequently entered into the GLMM algorithm.

The log-likelihood function from a sample of i.i.d. response variables for jth quarter over all combinations based on (3.1) is derived as

$$\ell(\xi_j|\{n_{ij}\}, \{\mu_{ij}\}) = \sum_{i=1}^{I} \left\{ n_{ij}\ln(\mu_{ij}) + n_{ij}\ln(\xi_j) - \left(n_{ij} + \frac{1}{\xi_j}\right)\ln(1+\xi_j\mu_{ij}) \right. \\ \left. + \ln\Gamma\left(n_{ij} + \xi_j^{-1}\right) - \ln\Gamma(n_{ij}+1) \right\} - I\ln\Gamma\left(\xi_j^{-1}\right), \tag{3.6}$$

where $j = 1, 2, \ldots, J$ and $\mu_{ij} = \exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T\boldsymbol{b}\right)$. During the M-H approximation process, $\boldsymbol{\beta}_j$ is generated from a multivariate normal distribution and $\boldsymbol{b}$ is generated under regression model conditioning on known $\boldsymbol{\beta}_j$ values at each iteration. Together with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, we can get the mean parameter μ_{ij}. Maximum likelihood estimation of ξ_j can then be obtained by unidimensional numerical maximization of $\ell(\xi_j|\{n_{ij}\}, \{\mu_{ij}\})$ given by (3.6). In each iteration, ξ_j is recalculated together with $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ from Gibbs sampling. All the newly generated parameter samples then provide a decision criteria in M-H algorithm.

3.3 Inference with Gibbs Sampler and Metropolis Step

3.3.1 Markov chain Monte Carlo for parameter estimations

In this subsection, we implement Markov chain Monte Carlo (MCMC) methods to explore and summarize posterior distributions using Bayesian statistics described in Section 3.2.1. Introduced by Metropolis et al. (1953) and Hastings (1970), MCMC has been a classical and general method for stochastic process simulations given probability density functions. It has been widely applied especially under the Bayesian algorithm (Gamerman and Lopes, 2006). Since it is not always feasible to find analytical expressions under the Bayes theorem for the posterior distribution of model parameters, Monte Carlo method (Metropolis and Ulam, 1949) has been brought up to estimate features of the posterior or predictive distribution of interest by using samples drawn from that distribution. One is able to simulate dependent samples from an irreducible Markov chain and treat stationary numerical approximations as an empirical distribution. Since M-H algorithm provides dependent chains, iteration samples require to be large enough in order to be independent.

In general, generating samples directly from a high dimensional joint distribution is unlikely possible. It is feasible to sample each parameter from the full conditional distribution via Gibbs sampler algorithm(Geman and Geman, 1984). As an indirect sampling approach, Gibbs sampling has become an increasingly popular statistical tool in both applied and theoretical research. Casella and George (1992) analytically establish its properties and provide insights on complicated cases. Smith and Roberts (1993) review the use of the Gibbs sampler for Bayesian computation and describe the implementation of MCMC simulation methods.

Based on the generalized parameterization scheme for our GLMM given by (3.1) and (3.3), $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{b}, \xi_j\}$ is a set of unknown parameters for jth quarter. The joint posterior distribution does not have a standard form and hence it is difficult to sample directly from it. Instead of getting a joint distribution of unknown parameters, we can construct a full conditional distribution $p(\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{b}, \xi_j | \boldsymbol{n}_1, \ldots, \boldsymbol{n}_J)$ by Gibbs sampler under M-H algorithm giving a MCMC approximation, where $\boldsymbol{n}_j = \{n_{1j}, \ldots, n_{Ij}\}$ represents a collection of data for the jth quarter. Iterated samplers from the full conditional distribution of each parameter generate a dependent sequence that converges to the joint conditional posterior distribution. The respective full conditional distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ rely only on $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$ as shown in (3.4) and (3.5) no matter what target distribution for Y_{ij} is chosen. Parameter $\boldsymbol{b}$ depends on the target distribution and is updated using given $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$ in each iteration. The remaining unknown dispersion parameter ξ_j is affected by the chosen NB-GLMM and its full conditional distribution, $f(n_{ij}|\mu_{ij}, \xi_j)$, can be obtained once the mean parameter μ_{ij} has been generated.

Given a set of starting values $\{\boldsymbol{\Sigma}^{(0)}, \boldsymbol{\beta}_1^{(0)}, ..., \boldsymbol{\beta}_J^{(0)}, \boldsymbol{b}^{(0)}\}$, the Gibbs sampler generates $(s+1)$th set of parameters $\{\boldsymbol{\theta}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)}, \xi_1^{(s+1)}, ..., \xi_J^{(s+1)}\}$ from $\{\boldsymbol{\theta}^{(s)}, \boldsymbol{\Sigma}^{(s)}, \boldsymbol{\beta}_1^{(s)}, ..., \boldsymbol{\beta}_J^{(s)}, \boldsymbol{b}^{(s)}\}$, $s = 0, 1, \ldots$. The logic of the Gibbs sampler updating algorithm can be described as follows.

1. Sample $\boldsymbol{\theta}^{(s+1)}$ from full conditional distribution (3.4):

   (a) compute $\boldsymbol{\mu}_J^{(s)}$ and $\Lambda_J^{(s)}$ from $\{\boldsymbol{\Sigma}^{(s)}, \boldsymbol{\beta}_1^{(s)}, ..., \boldsymbol{\beta}_J^{(s)}\}$, where

   $$\begin{aligned} \boldsymbol{\mu}_J^{(s)} &= (\Lambda_0^{-1} + J(\boldsymbol{\Sigma}^{(s)})^{-1})^{-1}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + J(\boldsymbol{\Sigma}^{(s)})^{-1}\bar{\boldsymbol{\beta}}^{(s)}), \\ \Lambda_J^{(s)} &= (\Lambda_0^{-1} + J(\boldsymbol{\Sigma}^{(s)})^{-1})^{-1}; \end{aligned}$$

   (b) sample $\boldsymbol{\theta}^{(s+1)} \sim \mathcal{N}\left(\boldsymbol{\mu}_J^{(s)}, \Lambda_J^{(s)}\right)$.

2. Sample $\boldsymbol{\Sigma}^{(s+1)}$ from full conditional distribution (3.5):

   (a) compute $\boldsymbol{S}_{\boldsymbol{\theta}}^{(s)}$ from $\{\boldsymbol{\theta}^{(s+1)}, \boldsymbol{\beta}_1^{(s)}, ..., \boldsymbol{\beta}_J^{(s)}\}$, where

   $$\boldsymbol{S}_{\boldsymbol{\theta}}^{(s)} = \sum_{j=1}^{J}(\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\theta}^{(s+1)})(\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\theta}^{(s+1)})^T;$$

(b) sample $\mathbf{\Sigma}^{(s+1)} \sim \mathcal{W}^{-1}\left(\nu_0 + J, \left[\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}}^{(s)}\right]^{-1}\right)$.

3. Obtain maximum likelihood estimate of $\boldsymbol{\xi}^{(s+1)} = \{\xi_1^{(s+1)}, ..., \xi_J^{(s+1)}\}$ from the conditional log-likelihood function (3.6), given $\{\boldsymbol{\beta}_1^{(s)}, ..., \boldsymbol{\beta}_J^{(s)}, \boldsymbol{b}^{(s)}\}$.

Such iterative algorithm constructs a dependent sequence of parameter values whose distribution converges to the target joint posterior distribution with a sufficiently large number of iterations. As seen from the algorithm, parameters $\{\boldsymbol{\theta}^{(s+1)}, \mathbf{\Sigma}^{(s+1)}, \boldsymbol{\xi}^{(s+1)}\}$ are sampled from the full conditional distributions or estimated from their log-likelihood functions; the set of parameter values are thus also samples from the joint distribution.

Given that $\boldsymbol{\theta}$ and $\mathbf{\Sigma}$ are estimated using conjugate prior distributions, their posterior distributions can be approximated with Gibbs sampler as described in Section 3.2.1. However, a conjugate prior distribution on $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J\}$ is not available due to high dimensions and full conditional distributions of the parameters do not have a standard form due to unknown sampling parameters. In this case, M-H algorithm can be a generic method to approximate the posterior distribution. M-H is named after Nicholas Metropolis (Metropolis et al., 1953) and W.K. Hastings (Hastings, 1970), which is a powerful Markov chain method to simulate multivariate distributions. Chib and Greenberg (1995) provide a tutorial introduction to the M-H algorithm and show examples on Gibbs sampler, a special case of the M-H algorithm. In our GLMM model, since the dominating density is not explicitly available, the M-H algorithm can be used under an acceptance-rejection scheme (Tierney, 1994). In acceptance-rejection step, we can generate candidates using Gibbs sampler from suitable generating density, and accept or reject observations from proposal distributions by implementing generation from a uniform distribution. Each step of the Gibbs sampler generates a proposal from full conditional distribution and then accept it. The Metropolis step generates proposals from population distribution and accepts them with some probability. M-H algorithm combines both approaches and allows arbitrary proposal distributions. Different from Metropolis's, acceptance ratio of Metropolis-Hastings is the probability of generating the current value from proposed to the probability of generating the proposed value.

For each $j \in \{1, ..., J\}$, Metropolis step for updating $\boldsymbol{\beta}_j^{(s)}$ by proposing a new value $\boldsymbol{\beta}_j^*$ from the multivariate normal distribution with the current mean value $\boldsymbol{\beta}_j^{(s)}$ and variance-covariance matrix $\mathbf{\Sigma}^{(s)}$ and accepting or rejecting it with appropriate probability described below. Then, $\boldsymbol{b}^{(s)}$ is to be updated by newly accepted $\{\boldsymbol{\beta}_1^{(s+1)}, ..., \boldsymbol{\beta}_J^{(s+1)}\}$.

1. Generate $\boldsymbol{\beta}_j^* \sim \mathcal{N}(\boldsymbol{\beta}_j^{(s)}, \mathbf{\Sigma}^{(s)})$.

2. Compute the acceptance ratio

$$r_j = \frac{\left[\prod_{i=1}^{I} f(n_{ij}|\mu_{ij}^*, \xi_j)\right] f(\boldsymbol{\beta}_j^*|\boldsymbol{\theta}^{(s)}, \mathbf{\Sigma}^{(s)})}{\left[\prod_{i=1}^{I} f(n_{ij}|\mu_{ij}^{(s)}, \xi_j)\right] f(\boldsymbol{\beta}_j^{(s)}|\boldsymbol{\theta}^{(s)}, \mathbf{\Sigma}^{(s)})},$$

where $\mu_{ij}^* = \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j^* + \boldsymbol{z}_{ij}^T \boldsymbol{b}^{(s)})$ and $\mu_{ij}^{(s)} = \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j^{(s)} + \boldsymbol{z}_{ij}^T \boldsymbol{b}^{(s)})$.

3. Sample $u \sim \text{uniform}(0, 1)$. Set $\boldsymbol{\beta}_j^{(s+1)}$ to $\boldsymbol{\beta}_j^*$ if $u < r$, or to $\boldsymbol{\beta}_j^{(s)}$ if $u > r$.

4. Update $\boldsymbol{b}^{(s+1)}$, given $\{\boldsymbol{\beta}_1^{(s+1)}, ..., \boldsymbol{\beta}_J^{(s+1)}, \boldsymbol{n}_1, ..., \boldsymbol{n}_J\}$, under our regression model given by (3.1) using the maximum likelihood algorithm.

In this way, the Gibbs sampler and Metropolis step described above are combined as an iterative algorithm to generate a Markov chain that can be used to approximate the joint posterior distribution of $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{b}, \boldsymbol{\xi}\}$. As iteration times go large enough so that the auto correlation effects are reduced, those sets of generated samples can be used to approximate the joint posterior distribution of all the parameters.

3.4 Analysis of Cyber Breach Chronology Data

Followed by empirical analysis presented in Section 3.2 and GLMM structure proposed in Section 3.3, we examine the manipulated PRC frequency dataset with unique subjective combinations. As mentioned in Section 2.2, the medical and non-medical portion (organization) of the data breach dataset are analyzed separately in our study. Since the only difference we treat between partitioned medical organization subdataset and non-medical organization subdataset is whether to include type of organizations as one of the covariates (we do not further partition medical organizations), we thus focus on the analysis of the non-medical portion of the PRC dataset with type of organizations factor in the rest of this chapter.

Quarterly counts of data breaches are modeled as a regression function of breach type, organization entity, entity location and overall quarterly average severity with specific identities under NB-GLMM. The effects due to potential trends overtime are also taken into consideration. We analyze in total 69 quarters (between years 2001 and 2018) of non-medical data breach incidents data in this section. Recall that in Section 2.3.1 levels of categorical covariates have been combined so there are 16 uniquely identified combinations (observations) within the non-medical subdataset. Therefore, among 69 investigated quarters, each quarter has 16 uniquely identified combinations that represent different cyber risk subjects, namely, unique type of data breaches, type of organizations and location of the entity that the breach incident occurs. Each combination can be treated as unique risk features/subjects corresponded to quarterly counts.

In order to detect the inner relationships between incident frequency and other features, a box plot is drawn in Figure 3.1 on frequency counts upon uniquely identified categorical level combinations for all the quarters under observation; it shows 16 boxes with each one representing the simplified distribution of 69 quarterly counts of that combination plotted upon uniquely identified level combinations. By examining these 16 distribution patterns of different combinations, we find that these count distributions differ significantly. For example, the 3rd and 8th combinations have higher log values of incident counts compared to other combinations, whereas the 12th combination has the lowest log median value of incident counts among all combinations.

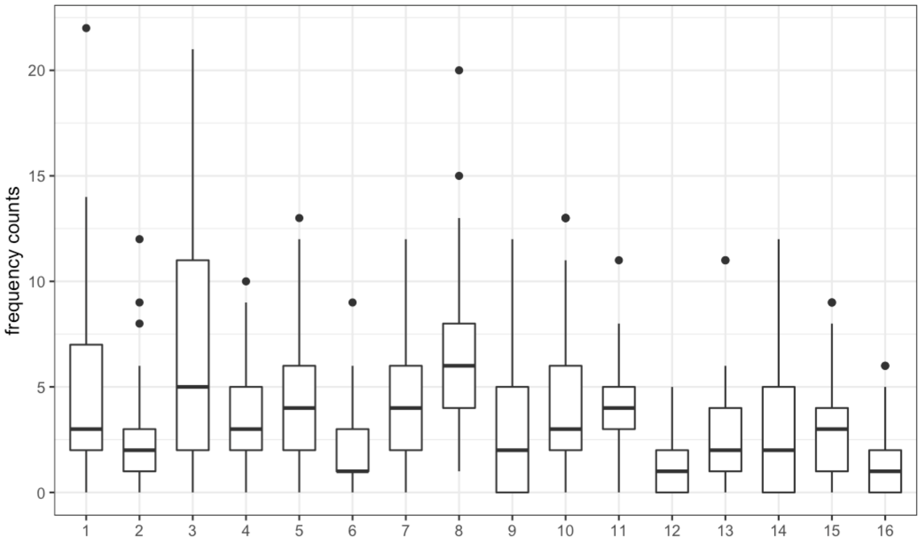
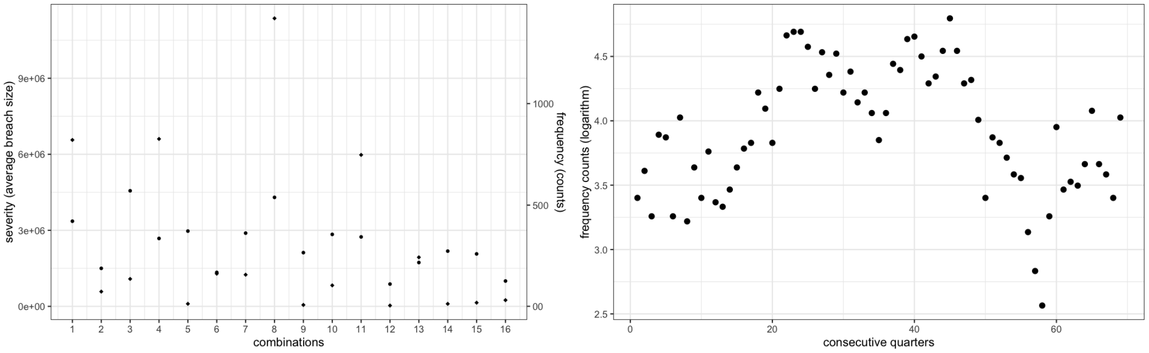


Figure 3.1: Quarterly frequency counts on individual categorical combinations

We also observe a correlation between quarterly counts and their corresponding average severity of combinations. Note that the severity here means the number of data breached caused by the data breach incident. It is observed that a quarter with high frequency counts often contains more incidents with a relatively large severity. Figure 3.2a is made up with scatter points of quarterly frequency (in rhombus) and corresponding average severity (in circle) showing that the dependence exists between counts and severity for most of combinations. This suggests that the average quarterly severity may be used as one of the covariates that impact on the quarterly counts of uniquely identified combinations.



(a) Scaled frequency and severity

(b) Polynomial time trend effect

Figure 3.2: Effects decomposition

Relationships between breach counts and classified characteristic combinations and severity dependency are significant among quarters. In this regard, we investigate the group specific variations by treating related covariate coefficients as multivariate normal random variables centering around a mean showed in (3.1). Coefficients can be decomposed into fixed effects representing overall magnitude for a given quarter and random effects representing the quarterly variation among quarters.

Besides within quarter fixed effects and among quarter random effects, there is potentially a time series relationship if we treat quarterly counts in a sequence timely manner. Figure 3.2b shows breach counts upon total 69 quarters in time sequence. The time series effect shows a polynomial trend which could be modeled by cubic polynomial time covariates. Cubic time trend is treated with only fixed effects with the remaining systematic noise being explained by random effects of quarterly variations.

Based on the findings showed above, we choose the following covariate manipulations for the generalized linear model used in (3.1):

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T \boldsymbol{b} = \sum_{l=1}^{6} x_{l,i} \beta_{l,j} + \sum_{k=1}^{3} z_{k,j} b_k, \tag{3.7}$$

where $\{x_{1,i}, x_{2,i}, x_{3,i}\}$ are the non-base level dummy variables of four regions under location covariate for the ijth count (ith combination in jth quarter), $\{x_{4,i}\}$ is the non-base level categories of type of breach for the ijth count, $\{x_{5,i}\}$ is the non-base level category of organization type and $\{x_{6,i}\}$ is the average severity of ith combination, $\{z_{1,j}, z_{2,j}, z_{3,j}\} = \{j, j^2, j^3\}$ are time, squared time and cubic time terms, measured in quarters. Here the effect of quarterly average severity is used by a numerical indicator to reveal the dependency between the frequency and severity. Details on the specific regions, types of data breaches and types of organizations can be found in Section 2.2. Regarding fix effects and random effects in (3.2), we assume random effects work on six factors which means $\boldsymbol{M}_j$ (for fixed effects) are different for different j's and $\boldsymbol{X}_j = \boldsymbol{X}$ (for random effects) is the same for all j's, and $\boldsymbol{u}_j$ follows a six-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariate matrix $\boldsymbol{\Sigma}$. Such a parameterization allows us not only to consider subject specific and group specific effects, but also to contain random effects on quarterly related factors other than time trends. In this subsection, the proposed NB-GLMM is used to analyze the quarterly data breach incidents recorded by PRC database using the M-G sampling algorithm under the Bayesian framework as described in Section 3.3.1. As discussed in Section 3.2.1, a multivariate normal distribution and an inverse-Wishart distribution are chosen as the prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, respectively. The starting values of hyper-parameters of both prior distributions are showed in Table 3.1.

The values for $\boldsymbol{\mu}_0$ are set as the mean of negative binomial regression coefficients without intercept, denoted by $\bar{\boldsymbol{\beta}}_{\text{GLM}}$, and for ν_0 is set as 8, which is the number of parameters $p = 6$ plus 2. Both $\boldsymbol{\Lambda}_0$ and $\boldsymbol{S}_0$ are set as the empirical variance-covariance matrix of regression

Parameter	Distribution	Starting Value
$\boldsymbol{\theta}$	$\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{\mathbf{0}})$	$\boldsymbol{\mu}_0 = \bar{\boldsymbol{\beta}}_{\text{GLM}}; \quad \boldsymbol{\Lambda}_{\mathbf{0}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\text{GLM}}}$
$\boldsymbol{\Sigma}$	$\mathcal{W}^{-1}\left(\nu_0, \boldsymbol{S}_0^{-1}\right)$	$\nu_0 = p+2; \quad \boldsymbol{S}_0 = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\text{GLM}}}$

Table 3.1: Simulation starting values

coefficients, denoted by $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\text{GLM}}}$. The starting values of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$, $\boldsymbol{b}$ and $\boldsymbol{\xi}$ used in the MCMC procedure are the negative binomial regression estimates. Total 69 Markov chains representing 69 quarters are generated at the same time in a matrix form in the model estimation process with $100,000$ iterations. In order to reduce autocorrelation, a thinning factor 10 is used. The first 200 iterations are discarded as burn-in samples and the remaining iterations are used for estimating the model parameters. A trace plot and autocorrelation function (ACF) are used to verify the proper convergence of simulation runs.

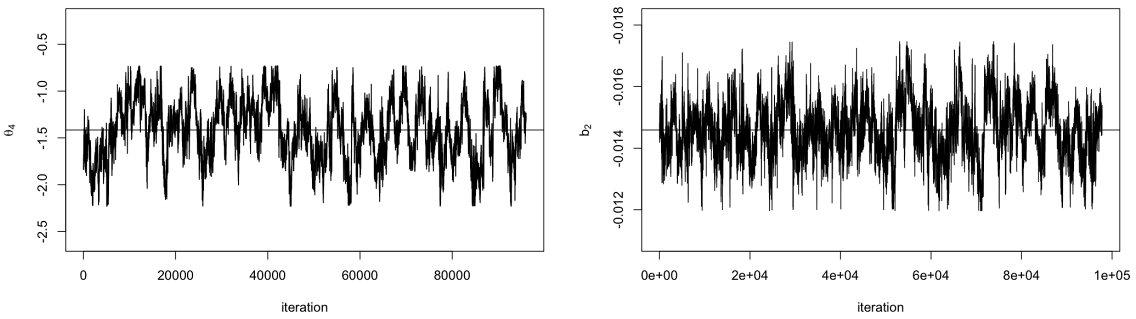
Table 3.2 displays the information about the posterior summary statistics of model parameters $\boldsymbol{\theta}$ and regression coefficients $\boldsymbol{b}$, including the posterior mean, standard error, and highest posterior density (HPD) intervals; the posterior means of the elements of the variance-covariance matrix $\boldsymbol{\Sigma}$ can be found in Appendix B. The results show that West region has the largest effects on number of counts per quarter. This may be because major tech companies are headquartered along the Pacific Coast where valuable gathered data are stored and shared over Internet. External breach type has a higher impact on breach frequency possibly because attackers tend to seek some types of benefit from breaching the victim's network. Business organizations receive more cyber breaches than non-business organization, which may be resulted from the reality that business organizations have various types of valuable information properties than non-business organizations do. As for the influence of average size, one unit increase in logarithm average severity causes a 0.8437-unit increase in breach counts on average.

Regressor	Symbol	Mean	Standard Error	95% HPD Interval	
South	θ_1	1.2536	0.0015	0.4053	2.2278
West	θ_2	2.2002	0.0011	1.4898	2.9617
Northeast	θ_3	0.7115	0.0011	0.0141	1.3812
Internal	θ_4	-1.4176	0.0011	-2.0852	-0.8232
Non-Business	θ_5	-0.2181	0.0011	-0.9858	0.3756
Ave-Size	θ_6	-0.1699	0.0001	-0.2322	-0.1103
Time^1	b_1	0.5892	9.0579×10^{-5}	0.5355	0.6997
Time^2	b_2	-1.4591×10^{-2}	2.7746×10^{-6}	-1.6347×10^{-2}	-1.2929×10^{-2}
Time^3	b_3	1.0075×10^{-4}	2.4653×10^{-8}	8.5920×10^{-5}	1.1628×10^{-4}

Note: Time^1, Time^2 and Time^3 represent the Time to the power $1, 2$ and 3, respectively.

Table 3.2: Posterior summary and interval statistics

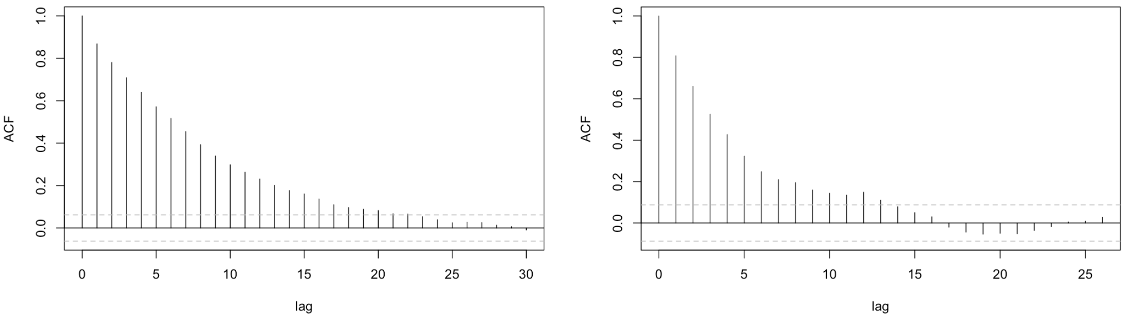
For each of the GLMM model parameters, MCMC generates a convergence diagnostic panel, which includes a trace plot, autocorrelation plot and a kernel density plot. We first assess if chains have run long enough for reliable estimations by monitoring convergence of iterative simulations (Brooks and Gelman, 1998), and then examine these diagnostic plots. Figures 3.3 and 3.4 show selected diagnostics for the slope coefficients θ_4 and b_2. Figures 3.3a and 3.3b are trace plots that show the number of iterations on the horizontal axis, plotted against the value of accepted coefficient of internal breach type θ_4 and b_2 on the vertical axis, respectively. Since there are no long term trends in these trace plots and the mixing is moving efficiently, we can affirm that the MCMC iteration converges. Figures 3.4a and 3.4b display the ACF values (Cowles and Carlin, 1996) of accepted coefficients θ_4 and b_2, respectively, at lag k on the vertical axis and k on the horizontal axis. Ideally, the autocorrelation at any lag should not be statistically significantly different from zero. It can be seen from the plot that the autocorrelations of θ_4 and b_2 are not significantly far from zero and the estimated autocorrelations are within the 95% confidence interval. These results support the conclusion that our MCMC iterations have converged.



(a) trace plot for θ_4

(b) trace plot for b_2

Figure 3.3: Trace plots



(a) autocorrelation plot for θ_4

(b) autocorrelation plot for b_2

Figure 3.4: Autocorrelation plots

3.5 Simulation Studies

We design a simulation study to verify the accuracy and effectiveness of the parameter estimations and the model predictability. The exploratory data analysis showed in this section should provide supports for the proposed NB-GLMM model. The simulation model is established in accordance with similar assumptions and design scheme of our analytical model. For demonstration purpose, this simulation study uses the same multivariate normal distribution estimated from Section 3.4. Given the sets of coefficients from multivariate normal distribution, we can generate target variable counts from generalized linear relationships. True values of model parameters are taken from Table 3.2 and Appendix B. According to the hierarchical requirements, we first draw 69 $\boldsymbol{\beta}_s$ from a 6-dimensional multivariate normal model with mean $\boldsymbol{\theta}$ and variance $\boldsymbol{\Sigma}$; together with posterior mean of $\boldsymbol{b}$, they consist 69 sets of independent quarter coefficients. Multiplying 69 sets of coefficients to the manipulated covariates using (3.7) leads to 69 logarithm mean of the negative binomial distribution. Combining those mean parameters with dispersion parameters we estimated previously, we generate 16 observations on uniquely identified combinations for each quarter, which results a total of 1104 observations. In this way we make sure that the simulated data follows the same patterns as experimental data. The new data set of 1104 testees is generated using the MCMC estimates obtained on the original dataset. Taking these observations as one dataset, we further generate 100 datasets following the same algorithm. Simulated datasets are then investigated under the same procedure as presented in Section 3.3. The estimated hyper-parameters are determined using MCMC and M-H methodologies, as well as maximum likelihood estimation under Bayesian framework. Here the MCMC analyses utilize the same prior distributions and the starting values are the same as obtained from the empirical estimation.

Regressor	Parameter	True Values	Estimated Mean	Relative Error
South	θ_1	1.2536	1.2018	-0.0413
West	θ_2	2.2002	2.2524	0.0237
Northeast	θ_3	0.7115	0.7429	0.0442
Int.	θ_4	-1.4176	-1.5368	0.0841
Non-Bus.	θ_5	-0.2181	-0.2335	0.0708
Ave-Size	θ_6	-0.1699	-0.1742	0.0255
Time^1	b_1	0.5892	0.5809	-0.0141
Time^2	b_2	-1.4591×10^{-2}	-1.4202×10^{-2}	-0.0267
Time^3	b_3	1.0075×10^{-4}	0.9913×10^{-4}	-0.0161

Note: Time^1, Time^2 and Time^3 represent the Time to the power $1, 2$ and 3, respectively.

Table 3.3: Simulation summary results

The estimated posterior means of coefficient parameters and the relative differences (errors) between the true and estimated values obtained under our modeling and estimation procedures are displayed in Table 3.3, where the relative error is calculated by dividing the difference of the estimated value and its corresponding true value by its true value (used for simulation). As seen from Table 3.3, differences between the true value and the estimated posterior means, illustrated by relative errors, are all relatively small. Relative error is a measure of the precision of an estimated population parameter. It quantifies the variability or dispersion of the sampling distribution of a statistic, most commonly the mean (Tibshirani and Efron, 1993). Having small standard errors imply that these estimated posteriors are all centered compactly around their true values. On the other hand, all the estimated results from our simulation study have over 99% confidence intervals where the true values fall into. All these imply that our estimation algorithm is effective and estimation results are satisfied in terms of their accuracy.

To examine the model predictability and its accuracy under our GLMM settings, we employ 5-fold cross-validation procedure to have an objective evaluation of the prediction performance. Cross-validation was first applied when evaluating the use of a linear regression equation for predicting a criterion variable (Mosier, 1951). It provides a more realistic estimate of model generalization error by repeating cross-validations based on the same dataset with large calibration/training samples and small validation/test samples. In particular, we randomly divide the dataset ten times into five folds; four of them are used to train the GLMM and remaining one is used to compare its predicted values and actual ones. The performance of the test datasets should be similar to that of the training datasets. Our purpose of conducting cross validations is to ensure that our model has not over-fitted the training dataset and that it performs well on the test dataset. In order to testify our GLMM prediction accuracy, we also fit our training dataset to Poisson and NB regression models, respectively. The root mean squared error (RMSE) metric is taken as a summary fit statistic, which can provide useful information for quantifying how well that our GLMM fits the dataset. A good performance with a relative low RMSE indicates that our proposed GLMM is fine-tuned. The RMSE values are calculated by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where n is the number of tested observations, y_i is the ith actual target value, and $\hat{y}_i$ is the ith predicted value based on trained model.

Table 3.4 gives summary fit statistics for Poisson regression, NB regression and NB-GLMM on training dataset and test dataset. We first compare training set RMSEs for model accuracy. The predicted accuracy of three models is compared under same training set measured by RMSE. The lowest training RMSE value of GLMM implies that it has the highest prediction level. We then compare GLMM RMSEs between the training set and the test set to test over-fitting. According to our cross validation results, the training set has a mean of 4.6384 RMSE which means that the average deviation between the 69 predicted quarterly counts and the actual quarterly ones is 4.6384.

A 4.8481 RMSE of the test dataset is close enough to that of the training dataset, which means that our model is not over-fitted. A higher RMSE of the test dataset is judged as an improvement in model fit when using the training dataset to build the model. Given the fact that two RMSEs do not have much difference, there is no evidence showing that our GLMM is over-fitted. These two relatively low values of RMSE also show that our model, GLMM, achieves the best model accuracy for frequency counts predictions among other tested models.

Partition	Training Set			Test Set
Model	Poisson	Negative Binomial	GLMM	GLMM
RMSE	5.1749	5.0516	4.6384	4.8481

Table 3.4: Summary fit statistics

3.6 Discussion

In this section, we discuss the potential applications and practical implications of our modeling results in cyber risk mitigation and management. We have proposed a NB-GLMM with group-specific fixed effects and among group random effects on some featured variables including the type of breached, type of organizations and their geographical location and associated average severity caused by data breaches under these uniquely identified features. We also consider the impact of the trend over time on the breach frequencies. In general, this study can increase the awareness that it is important to analyze the growth trends of cyber incidents frequency among sub-characteristic groups. We discuss below the impact of our modeling and predictive analytic approaches in relation to cyber risks from both the perspective of the organization (potential insured) and the insurance company (insurer), as well as other important stakeholders such as corporate information technology (IT) and data security officers, and data scientists.

From the perspective of organizations, our results provide quantitative insights to organizations with different entity types and locations, which encourages firms to adopt new techniques and technologies in managing risks with respect to the cyber-related risks they are facing. Gordon and Loeb (2002) present an economic model that can be used to determine the optimal amount to invest to protect a given set of information. The model takes into consideration the vulnerability of the information to a security breach and the potential loss it may cause. Given a company's physical and geographical characteristics, our NB-GLMM model is able to predict their estimated quarterly data breach frequencies; by combining the severity model that we propose in Chapter 4 for the data breaches, the firms can determine whether to accept the risk or to seek out risk transformation in order to mitigate risks. Mazzoccoli and Naldi (2020) propose an integrated cyber risk management strategy that combines insurance and security investments, and investigate whether it can be used to reduce overall security expenses. The optimal investment for their proposed mixed strategy is derived under several insurance policies. This type of risk management strategies could also include the consideration of the risk over a specified time horizon; our model can provide an effective predictive guidance for managing cyber risks with respect to data breach incidents occurred within a quarterly time interval. The organizations could act based on our findings when they put cyber risk management into practice.

In some cases, managing cyber risks through internal controls would be impractical or too costly especially when organizations are facing high frequency of breach incidents. Consequently, organizations may seek insurance coverage as alternative means to transfer their cyber related risks. Reducing cyber risk exposures by purchasing insurance also takes advantage of reducing the capital that must be allocated to the cyber risk management. In general, cyber insurance combined with adequate security system investments should allow organizations to better manage their cyber-related risks. Young et al. (2016) present a framework that incorporates insurance industry operating principles to support quantitative estimates of cyber-related risks and the implementation of mitigation strategies.

From the perspective of insurance companies, besides those incentives from organizations to increase cyber insurance purchases, our results also encourage insurance companies to consider how much premiums they should collect because they expect to be paid adequately to accept the risk. The current pricing of cyber insurance products is based on expert models rather than on historical data. An empirical approach to identifying and evaluating potential exposure measure is important but challenging due to the current scarcity of reliable, representative and publicly available loss experience for cyber insurance. This study avoids this limitation by illustrating how to utilize available full exposure data to get a quantitative idea of cyber premium pricing. We present a methodology to rigorously classify different risk levels of insureds. Our modeling results can ease one of the problems that cyber risk insurers face, the disparity in premiums with respect to different characteristic groups, by forecasting the loss frequency on different characteristic segmentations. Geographical area is one of the most well-established and widely-used rating variables, whereas business type is considered as one of the primary drivers of cyber claims experience.

Ideally, the cyber insurance rating system should consider various rate components, such as business type and geographic location in our model, when calculating the overall premium charged for cyber risks. The portion of the total premium that varies by risk characteristics, shown as a function of the base rate and rate differentials, is referred to as a variable premium (Werner and Modlin, 2010). Our work can be directly applied in setting variable premium factors by using posterior frequency distributions upon different risk characteristic segments. For example, the premium P under the standard deviation premium principle (Tse, 2009) for pricing variable premium is given by

$$P = \mathrm{E}[S] + \vartheta\sqrt{\mathrm{Var}(S)},$$

where S is the aggregated total loss, and ϑ is the loading factor. To calculate the premium rate P in this case, the first two moments of the distribution of S need to be determined. When ϑ is set to be zero, $P = \mathrm{E}[S]$ is the base (pure) premium, and together with a given industry risk loading ϑ, differential premium factors based on risk characteristics can be determined as a function of the loading factor and the standard deviation of the total loss amounts.

In addition to the idea of defining risk classes, this study illustrates how to work with current available data and update the model components and parameters by collected cyber related data over time. Our model decomposes risk effects on cyber breach frequencies into fixed effects and random effects based on classified characteristics, average severity and non-linear time trend effects. Bayesian statistics are particularly useful in simulating from the posterior distribution of the number of incidents (claims) in a future quarterly based time period given risk characteristics. Due to the nature of Bayesian methodology, some of the assumptions, such as the polynomial time trend, and parameters choices might be updated in the future once suitable data is available. Moreover, individual features of the model can be refined or replaced to incorporate properties of given internal datasets without changing the overall model structure. The updates and modifications enable our model to be a precise predictor for data breach frequencies.

This study develops a statistical model for cyber breach frequencies that considers not only characteristics such as risk profile, location and industry, but also average loss sizes and time effects. It provides an effective and comprehensive modeling approach for predictive analytics due to the consideration of dependent and correlated risk aspects. We believe that our study makes an important and novel contribution to the actuarial literature in the sense that our NB-GLMM for cyber breach frequencies considers risk category, company census, severity dependence and time trend effects together in quantifying and predicting quarterly number of data breach incidents, a fundamental quantity for appropriately setting the manual rates.

The study of cyber risks is important for insurance companies in mitigating and managing their risks given that the functioning of the insurance business is a complex process. In this view, our study is of practical value for insurance companies, since the consideration of the most dangerous risks for each business entity will allow forming a relevant information security for the company. Enterprises need to take several measures in dealing with cyber risks: operations based on statistical modeling in actuarial analysis process, ensuring the balance and adequacy of tariffs in pricing process and adjusting premium rates in insurance marketing. Our research results can be used as a differential indicator on different organization types and geographical locations. In addition, our study can also be useful for data security officers and scientists, and other potential corporate stakeholders for them to better understand the impact of the cyber risks to business operations.

Another important aspect of this study is the use of the publicly available PRC data on developing actuarial approaches to quantify cyber loss frequencies. However, the quality of available data and whether the data represents well cyber risks in general also lead to a limitation of our study. The fact that firms do not reveal details concerning security breaches reduces data accuracy, and not voluntarily reporting cyber breaches leads to data inadequacy. Moreover, Privacy Rights Clearinghouse (PRC) has stopped updating latest breach incidents since 2019, which causes data inconsistency in a time trend manner. The availability of high-quality data such as policy or claim database in the future would open up new research opportunities. Our model is subjective and can be modified to accommodate the features of new dataset and the purpose of prediction.

Despite the limitations, the proposed NB-GLMM makes a notable methodological contribution to the cyber insurance area as it provides a theoretically sound modeling perspective in frequency quantification, and provides a practical and statistical framework and approach for practitioners to customize and update based on their predictive needs. In Chapter 4, we analyze zero-inflated heavy tailed loss amounts (the number of data breached due to breach incidents and their corresponding monetary losses incurred) using finite mixture model and extend the analysis using the extreme value theory. Together with NB-GLMM frequency predictive model, we can simulate aggregate full insurance losses with given characteristics. Moreover, we will use a numerical approach to test predicted overall quarterly aggregate claim amounts under different factor combinations in order to make characterization of premiums. For instance, pure technical insurance premiums can be expressed as a VaR or TVaR metric and computed from the loss distribution of each risk category. Lastly, this two-part severity-frequency actuarial quantification method seeks to overcome some of above-mentioned data limitations such as inadequacy and inconsistency.

Chapter 4

Zero-inflated mixture and composite regression model for Cyber Loss Severity Analysis

After we discuss the loss frequency modeling in Chapter 3, we present in this chapter zero-inflated mixture and composite regression (Zi-MCR) model for loss severity and discuss their application in cyber risk analysis. Section 4.1 reviews the definition of splicing models and finite mixture models, which are the composition of mixture and composite regression model (MCR), and propose our unique MCR model adjusted by zero-inflated component based on the distinct feature of the dataset we study. Followed by Section 4.2, we introduce the expectation-maximization (EM) algorithm to be used to estimate coefficients and model parameters including E-step, M-step and starting values. We then present details on how to fit the PRC dataset to the models we propose as well as on how to assess the goodness of fit of a model in Section 4.3. Finally, we discuss the model application in terms of loss severity from the insurers' perspective in Section 4.4.

4.1 Notation and Modeling

One of the professional responsibilities of actuaries is to study loss distributions (patterns) based on the data collected. As seen from the empirical data analysis in Chapter 2, the severity distribution (pattern) of cyber loss records possesses point masses of zero, features over-dispersion and a relatively long tail nature, and shows different patterns of loss amounts (the number of data breached) for medical and non-medical organizations, which can hardly be fitted by a single analytic and parametric distribution. Our dataset also allows us to examine individual risk characteristics via regression predictors, such as breach type, business type, and location. Based on these characteristics, we propose a finite mixture model with three components integrated with a GLM framework to analyze the severity of cyber losses.

4.1.1 Splicing models

The distribution of loss variables, such as bodily injury costs and cyber losses, often features long tails. Consequently, when modelling claim sizes to set premiums, calculating risk measures, and determining capital requirements for solvency regulations, it is frequently necessary for the actuarial analytic domain to obtain a global fit for loss/risk distributions. In the literature, a splicing model is also called a composite model, in which multiple light-tailed distributions for the body and a heavy-tailed distribution for the tail are combined. The general density form of an m-component spliced distribution can be expressed as

$$f(y) = \begin{cases} p_1 f_1(y) & y \in C_1, \\ p_2 f_2(y) & y \in C_2, \\ \vdots & \\ p_m f_m(y) & y \in C_m, \end{cases} \tag{4.1}$$

where f_i, for $i = 1, 2, \ldots, m$, are legitimate density functions defined on the respective mutually exclusive and sequentially ordered intervals $C_1, C_2, \ldots, C_m$ with corresponding positive weights $p_1, p_2, \ldots, p_m$ that add up to one, i.e., $\sum_{i=1}^m p_i = 1$. In this regard, the density function f given by (4.1) and its corresponding cumulative distribution function F can be written, for $y \in \bigcup_{i=1}^m C_i$, in a compact form as

$$f(y) = \sum_{i=1}^m I_{C_i}(y) p_i f_i(y), \qquad F(y) = \sum_{i=1}^m I_{C_i}(y) \left(\sum_{j=1}^{i-1} p_j + p_i F_i(y) \right),$$

where I is an indicator function with $I_{C_i}(y) = 1$, if $y \in C_i$, otherwise 0, F_i is the corresponding cumulative distribution function of f_i in the interval C_i.

Based on the empirical analysis results shown in Table 2.5, we consider a spliced distribution with three components: the first component contains zeros, the second component models the middle segment of the amount of loss data, and the third component models the tail segment. Let Y denote the random variable that represents the jth loss amount, c is a non-zero loss threshold, and then the pdf of Y can be expressed as

$$f(y|\boldsymbol{\zeta}) = \begin{cases} p_1(\boldsymbol{\zeta}) & y = 0, \\ p_2(\boldsymbol{\zeta})\frac{f_1(y;\boldsymbol{\zeta}_1)}{F_1(c;\boldsymbol{\zeta}_1)-F_1(0^+;\boldsymbol{\zeta}_1)} & y \in (0, c], \\ [1 - p_1(\boldsymbol{\zeta}) - p_2(\boldsymbol{\zeta})]\frac{f_2(y;\boldsymbol{\zeta}_2)}{1-F_2(c;\boldsymbol{\zeta}_2)} & y \in (c, \infty), \end{cases}$$

where f_1 and f_2 are two density functions with cdf F_1 and F_2, defined on $(0, c]$ and (c, ∞), respectively, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1^T, \boldsymbol{\zeta}_2^T)^T$ is a set of parameter vector associated with the distributions of the components F_1 and F_2. The splicing weights p_1 and p_2 are functions of $\boldsymbol{\zeta}$ and can be estimated from the proportions of points equal to 0, and falling in the intervals $(0, c]$ and (c, ∞). The threshold c is a parameter to be estimated from the data which is investigated in Section 4.2.3. The remaining unknown parameters p_1, p_2 and $\boldsymbol{\zeta}$ can be estimated using the maximum likelihood estimation (MLE) method by maximizing the log-likelihood function based on observations $y_1, y_2, \ldots, y_n$, which is given by

$$\begin{aligned} &\log \mathcal{L}(\boldsymbol{\zeta}) \\ &\quad = \log(p_1(\boldsymbol{\zeta})) \sum_{j=1}^{n} I_{\{0\}}(y_j) \\ &\qquad + \sum_{j=1}^{n} I_{(0^+,c]}(y_j) \Big[\log(p_2(y_j;\boldsymbol{\zeta})) + \log f_1(y_j;\boldsymbol{\zeta}_1) - \log(F_1(c;\boldsymbol{\zeta}_1) - F_1(0^+;\boldsymbol{\zeta}_1))\Big] \\ &\qquad + \sum_{j=1}^{n} I_{(c,\infty)(y_j)} \left[\log(1 - p_1(\boldsymbol{\zeta}) - p_2(\boldsymbol{\zeta})) + \log f_2(y_j;\boldsymbol{\zeta}_2) - \log(1 - F_2(c;\boldsymbol{\zeta}_2))\right]. \end{aligned}$$

4.1.2 Finite mixture models

Due to the adaptability in utilizing high-dimensional features, coping with population heterogeneity, and achieving multiple interrelated goals, mixture distributions have gained popularity in recent years. Peel and MacLahlan (2000) provide a thorough discussion of using the EM algorithm to find maximizers of MLE and the selection of the number of components in finite mixture models. Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample of size n, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the observed value of random vector. Suppose that Y_j follows a finite mixture distribution with density function f on $\mathbb{R}$, which can be written in the form[1]

$$f_M(y_j) = \sum_{i=1}^{g} \pi_i f_i(y_j), \tag{4.2}$$

where for $i = 1, 2, \ldots, g$, f_i is a density function on $\mathbb{R}$ and π_i is a non-negative quantity such that $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{g} \pi_i = 1$. The quantities $\pi_1, \pi_2, \ldots, \pi_g$ are the mixing proportions or weights, and $f_1, f_2, \ldots, f_g$ are called the component densities of the mixture. We call density (4.2) as a g-component finite mixture density function and its corresponding distribution F_M as a g-component finite mixture distribution function.

In order to well interpret the mixture models, let Z_j be a categorical random variable taking values in $\{1, 2, \ldots, g\}$ with probabilities $\pi_1, \pi_2, \ldots, \pi_g$, respectively, and suppose that the conditional density of $\boldsymbol{Y}_j$ given $Z_j = i$ is $f_i(\boldsymbol{y}_j)$ $(i = 1, 2, \ldots, g)$ whereas the unconditional/marginal density is $f(\boldsymbol{y}_j)$. Other than that, it is convenient to work with a g-dimensional component label vector $\boldsymbol{Z}_j$ in place of the single categorical variable Z_j, where the ith element of $\boldsymbol{Z}_j$, $Z_{ij} = (\boldsymbol{Z}_j)_i$, is defined to be one or zero, according to whether the component of origin of $\boldsymbol{Y}_j$ in the mixture is equal to i or not $(i = 1, 2, \ldots, g)$. Thus in such setting, this categorical random vector $\boldsymbol{Z}_j$ can be viewed as following a multinomial distribution with probabilities $\pi_1, \pi_2, \ldots, \pi_g$; that is,

$$\mathrm{P}\{\boldsymbol{Z}_j = \boldsymbol{z}_j\} = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \cdots \pi_g^{z_{gj}}, \tag{4.3}$$

according to a multinomial distribution consisting of one draw on g categories with probabilities $\pi_1, \pi_2, \ldots, \pi_g$. We write

$$\boldsymbol{Z}_j \sim \mathrm{Mult}_g(1, \boldsymbol{\pi}), \qquad \boldsymbol{Z}_j \in \{0, 1\}^g,$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_g)^T$. In the interpretation of a mixture model, Y_j is drawn from a population with g groups, $G_1, G_2, \ldots, G_g$, with proportions $\pi_1, \pi_2, \ldots, \pi_g$, where the density of Y_j in group G_i is $f_i(y_j)$. The component-indicator variables z_{ij} are used in finding optimizers under ML estimation via the EM algorithm to be discussed in Section 4.2.

Generally, the components can be any exponential family distribution (Hasselblad, 1969); observations are available from a population known to be a mixture of K subpopulations. In our study, each subpopulation is not necessarily assumed to have the same type of distribution, which is one of the most significant departures from previous research. For a single observation y_j, the probability density of the exponential family can be ex-

[1] In this formulation of the mixture model, the number of components g is considered fixed. In many applications, the value of g is unknown and inferred from the available data, along with the mixing proportions and the parameters in the specified forms of the component densities.

pressed as

$$f(y_j;\theta_j,\phi) = \exp\left\{\frac{y_j\theta_j - b(\theta_j)}{\phi} + c(y_j,\phi)\right\}, \tag{4.4}$$

where θ_j is a natural parameter or canonical parameter, ϕ is the dispersion parameter or scale parameter[2], $b(\theta_j)$ is a known function of θ_j called cumulant function and $c(y_j,\phi)$ is a normalizing function, ensuring that (4.4) is a probability function. The mean and variance of exponential family distributions can be expressed by $b(\theta_j)$ (assuming twice differentiable) as follows:

$$E(Y_j) = \mu_j = b'(\theta_j), \qquad \mathrm{Var}(Y_j) = \phi\, b''(\theta_j) = \phi V(\mu_j), \tag{4.5}$$

where the variance of Y_j is the product of two terms, the dispersion parameter ϕ and the variance function $V(\mu_j) = b''(\theta_j)$, which is usually written in the following form

$$V(\mu_j) = \frac{\partial \mu_j}{\partial \theta_j}.$$

Considerations are given to the family of mixtures of GLMs, because many applications of non-normal mixtures involve components from the exponential family. The GLM is a statistical framework for unifying several significant exponential family models (Nelder and Wedderburn, 1972). Under this framework, it is permissible for the mixing proportions and the component distributions to depend on some associated covariates. In a GLM setting, it is assumed that

$$\eta_j = h(\mu_j) = \boldsymbol{x}_j^T\boldsymbol{\beta}, \tag{4.6}$$

where $h(\cdot)$ is a monotonic function known as the link function, η_j is the linear predictor, and μ_j, the mean of an exponential family distribution $f(y_j;\theta_j,\phi)$, is a known function of the canonical parameter θ_j described in (4.5), $g(\cdot)$ is a known link function that connects the distribution mean and the linear combination of explanatory variables together.

With the methodologies described above, we propose a finite mixture regression model where mixture components can be from a same type or different types of parametric family. Our model employs candidate distributions such as Gamma, Log Normal, Inverse Gaussian, and Weibull from the exponential family because the loss or severity is typically modelled using continuous random variables.

4.1.3 Zero-inflated mixture and composite regression models (Zi-MCR)

In this subsection, in order to provide a clear understanding of our combined finite mixtures and splicing model, we first introduce our model under a general splicing framework with three parts spliced densities jointing with weighting probabilities, followed by a de-

[2] When ϕ is known, the distribution of Y_j is one-parameter canonical exponential family member. When ϕ is unknown, it is often a nuisance parameter and then it is estimated by the method of moments. In most of GLM theory, the role of ϕ is often treated as an unknown constant but not as a parameter. (Yee, 2015)

tailed description of finite mixture section of the spliced density part for the moderate loss amounts.

Let Y be a non-negative claim severity random variable, and let $\boldsymbol{x} \in \mathbb{R}^p$ be the vector of covariate information. The density of the zero-inflated mixture composite regression model written in a spliced form with zero and two densities f_M and f_T and their corresponding cumulative distribution functions (CDFs) F_M and F_T is given by

$$\begin{aligned} & f_Y(y_j; \boldsymbol{\alpha}, \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \kappa, \boldsymbol{x_j}) \\ & = p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}) \mathbf{1}\{y_j = 0\} \\ & \quad + p_2(\boldsymbol{\alpha}, \boldsymbol{x_j}) \frac{f_M(y_j; \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{x_j})}{F_M(c; \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{x_j}) - F_M(0^+; \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{x_j})} \mathbf{1}\{0 < y_j \leq c\} \\ & \quad + [1 - p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}) - p_2(\boldsymbol{\alpha}, \boldsymbol{x_j})] \frac{f_T(y_j; \boldsymbol{\gamma}, \kappa, \boldsymbol{x_j})}{1 - F_T(c; \boldsymbol{\gamma}, \kappa, \boldsymbol{x_j})} \mathbf{1}\{y_j > c\}, \end{aligned} \tag{4.7}$$

where $\{p_1, p_2\} \in (0, 1)$ are the splicing weights, c is the splicing point which is the threshold separating the moderate and extreme loss values, $\boldsymbol{\alpha}$ is covariate coefficients of zero-inflated weight, $\mathcal{W}$, $\mathcal{B}$ and $\boldsymbol{\phi}$ are parameter vectors of the density of body f_M which is a finite mixture model, and $\boldsymbol{\gamma}$ and κ are parameters of the density of tail f_T.

In this study, the finite mixture distribution f_M is the density of positively defined continuous distributions with upper truncation at the threshold loss level c, given by

$$f_M(y_j; \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{x_j}) = \sum_{i=1}^{g} \pi_{ij}(\mathcal{W}, \boldsymbol{x_j}) f_i \left(y_j; \exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}_i \right), \phi_i \right) \tag{4.8}$$

where $\mathcal{B} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_g^T)^T$ is regression coefficients, and $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_g)^T$ is a vector of fixed dispersion parameters of g distribution components from the exponential family. The parameter π_{ij} is the mixing proportion of the ith function and jth observation which is a function of $\boldsymbol{x}_j$ and commonly modeled by logistic distributions

$$\pi_{ij} = \pi_i(\mathcal{W}, \boldsymbol{x}_j) = \frac{\exp\left(\boldsymbol{x}_j^T \boldsymbol{\omega}_i \right)}{1 + \sum_{h=1}^{g-1} \exp\left(\boldsymbol{x}_j^T \boldsymbol{\omega}_h \right)}, \tag{4.9}$$

where $\mathcal{W} = (\boldsymbol{\omega}_1^T, \ldots, \boldsymbol{\omega}_{g-1}^T, \boldsymbol{\omega}_g^T)^T$, with $\boldsymbol{\omega}_g = \mathbf{0}$, contains the logistic regression coefficients. Lastly, f_T is the tail density function from the exponential family with heavy-tailed performance, given by

$$\begin{aligned} f_T(y_j; \boldsymbol{\gamma}, \kappa, \boldsymbol{x_j}) &= f_T \left(y_j; \exp\left(\boldsymbol{x}_j^T \boldsymbol{\gamma} \right), \kappa \right) \\ &= \exp\left\{ \frac{y_j \theta_j - b(\theta_j)}{\kappa} + c(y_j, \kappa) \right\}, \end{aligned}$$

where $\exp\left(\boldsymbol{x}_j^T \boldsymbol{\gamma}_j \right) = \theta_j$, the canonical parameter in (4.4), and κ is the dispersion parameter.

4.2 Parameter Inferences by EM Algorithm

Let $\boldsymbol{\Psi}$ denote the set of vectors representing all the unknown parameters in (4.7) that need to be estimated, namely,

$$\boldsymbol{\Psi} = (\boldsymbol{\alpha}^T, \mathcal{W}^T, \mathcal{B}^T, \boldsymbol{\gamma}^T)^T. \tag{4.10}$$

The density of our spliced mixture regression model (4.7) of the jth response variable Y_j, for $j = 1, 2, \ldots, n$, can then be written as follows:

$$\begin{aligned} f(y_j; \boldsymbol{\Psi}, \boldsymbol{x}_j) = p_{1_j}\mathbf{1}\{y_j = 0\} + p_{2_j}\frac{\sum_{i=1}^{g} \pi_{ij} f_i(y_j; \boldsymbol{\beta}_{ij}, \phi_i, \boldsymbol{x}_j)}{F_M(c; \boldsymbol{\Psi}, \boldsymbol{x}_j) - F_M(0^+; \boldsymbol{\Psi}, \boldsymbol{x}_j)}\mathbf{1}\{0 < y_j \leq c\} \\ + (1 - p_{1_j} - p_{2_j})\frac{f_T(y_j; \boldsymbol{\gamma}_j, \kappa, , \boldsymbol{x}_j)}{1 - F_T(c; \boldsymbol{\gamma}_j, \kappa, , \boldsymbol{x}_j)}\mathbf{1}\{y_j > c\}. \end{aligned}$$

where, for simplicity of notation, $p_{1_j} = p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})$, $p_{2_j} = p_2(\boldsymbol{\alpha}, \boldsymbol{x}_j)$ and $\pi_{ij} = \pi_i(\mathcal{W}, \boldsymbol{x}_j)$ represents the mixing proportion in (4.12). In this way, the log likelihood for $\boldsymbol{\Psi}$ can be formed as

$$\begin{aligned} \log\mathcal{L}(\boldsymbol{\Psi}) = & \sum_{j=1}^{n} \log(p_{1_j})\mathbf{1}\{y_j = 0\} + \sum_{j=1}^{n}\Bigg\{\log(p_{2_j}) + \log\left(\sum_{i=1}^{g} \pi_{ij} f_i(y_j; \boldsymbol{\beta}_{ij}, \phi_i, \boldsymbol{x}_j)\right) \\ & - \log[F_M(c; \boldsymbol{\Psi}, \boldsymbol{x}_j) - F_M(0^+; \boldsymbol{\Psi}, \boldsymbol{x}_j)]\Bigg\}\mathbf{1}\{0 < y_j \leq c\} \\ & + \sum_{j=1}^{n}\Bigg\{\log(1 - p_{1_j} - p_{2_j}) + \log f_T(y_j; \boldsymbol{\gamma}_j, \kappa, \boldsymbol{x}_j) \\ & - \log[1 - F_T(c; \boldsymbol{\gamma}_j, \kappa, \boldsymbol{x}_j)]\Bigg\}\mathbf{1}\{y_j > c\}. \end{aligned}$$

The EM algorithm (Dempster et al., 1977) can be applied to obtain the MLE of $\boldsymbol{\Psi}$ in this spliced mixture regression model. The complete-data log likelihood is given by

$$\begin{aligned} \log\mathcal{L}_c(\boldsymbol{\Psi}) = & \sum_{j=1}^{n} \log(p_{1_j})\mathbf{1}\{y_j = 0\} \\ & + \sum_{j=1}^{n}\Bigg\{\log(p_{2_j}) + \sum_{i=1}^{g} z_{ij}[\log(\pi_{ij}) + \log f_i(y_j; \boldsymbol{\beta}_{ij}, \phi_i)] \\ & - \log[F_M(c; \boldsymbol{\Psi}) - F_M(0^+; \boldsymbol{\Psi})]\Bigg\}\mathbf{1}\{0 < y_j \leq c\} \\ & + \sum_{j=1}^{n}\Bigg\{\log(1 - p_{1_j} - p_{2_j}) + \log f_T(y_j; \boldsymbol{\gamma}_j, \kappa, \boldsymbol{x}_j)] - \log(1 - F_T(c; \boldsymbol{\gamma}_j, \kappa, \boldsymbol{x}_j))\Bigg\}\mathbf{1}\{y_j > c\} \end{aligned} \tag{4.11}$$

where z_{ij} denotes the component-indicator variables as defined in (4.3). Note that the composite probabilities p_1 and p_2 are to be estimated outside of E-M steps using proportions of zero for each covariate combination and presented in Section 4.2.3.

4.2.1 E-step

The EM algorithm is applied to this problem by treating the z_{ij} as missing data. E (for expectation) and M (for maximization) are the two iterative steps. Given an observed data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$, we take the conditional expectation of the complete-data log likelihood (4.11) using the current fit for $\boldsymbol{\Psi}$. We consider $\boldsymbol{\Psi}^{(0)}$ as an initial value of the iterative computation. The E-step computes the conditional expectation of $\log\mathcal{L}_c(\boldsymbol{\Psi})$ given $\boldsymbol{y}$ using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$ on the first EM algorithm iteration, that is,

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = \mathrm{E}_{\boldsymbol{\Psi}^{(0)}}[\log\mathcal{L}_c(\boldsymbol{\Psi})|\boldsymbol{y}],$$

The expectation operator E has the subscript $\boldsymbol{\Psi}^{(0)}$ to explicitly convey that this expectation is being effected using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$. In this manner, the E-step calculates $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ on the $(k+1)$th iteration, where $\boldsymbol{\Psi}^{(k)}$ is the value of $\boldsymbol{\Psi}$ after kth iteration. The E-step requires the calculation of the current conditional expectation of Z_{ij} given the observed data $\boldsymbol{y}$, which can be calculated as

$$\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}) = \mathrm{P}_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} = 1|\boldsymbol{y}\} = \tau_{ij}(y_j; \boldsymbol{\Psi}^{(k)}), \tag{4.12}$$

where for $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, n$,

$$\begin{aligned}\tau_{ij}(y_j; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_j) &= \pi_{ij}^{(k)} \frac{f_i(y_j; \boldsymbol{\beta}_{ij}^{(k)}, \phi_i, \boldsymbol{x}_j)}{f_M(y_j; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_j)} \\ &= \pi_{ij}^{(k)} \frac{f_i(y_j; \boldsymbol{\beta}_{ij}^{(k)}, \phi_i, \boldsymbol{x}_j)}{\sum_{h=1}^{g} \pi_{hj}^{(k)} f_h(y_j; \boldsymbol{\beta}_{hj}^{(k)}, \phi_h, \boldsymbol{x}_j)},\end{aligned}$$

where $\pi_{ij}^{(k)} = \pi_i(\mathcal{W}^{(k)}, \boldsymbol{x}_j)$ based on (4.9). The quantity $\tau_{ij}(y_j; \boldsymbol{\Psi}^{(k)})$ is the posterior probability that the jth member of the sample with observed value y_j belongs to the ith component of the mixture. Taking the conditional expectation of (4.11) using (4.12) gives that

$$
\begin{aligned}
Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) =& \sum_{j=1}^{n}\log(p_{1_j})\mathbf{1}\{y_j=0\} \\
&+\sum_{j=1}^{n}\Bigg\{\log(p_{2_j})+\sum_{i=1}^{g}\tau_{ij}(y_j;\boldsymbol{\Psi}^{(k)},\boldsymbol{x}_j)[\log(\pi_{ij})+\log f_i(y_j;\boldsymbol{\beta}_{ij}^{(k)},\phi_i,\boldsymbol{x}_j)] \\
&-\log[F_M(c;\boldsymbol{\Psi}^{(k)},\boldsymbol{x}_j)-F_M(0^+;\boldsymbol{\Psi}^{(k)},\boldsymbol{x}_j)]\Bigg\}\mathbf{1}\{0<y_j\le c\} \\
&+\sum_{j=1}^{n}\Bigg\{\log(1-p_{1_j}-p_{2_j})+\log f_T(y_j;\boldsymbol{\gamma}_j^{(k)},\kappa,\boldsymbol{x}_j) \\
&\qquad -\log(1-F_T(c;\boldsymbol{\gamma}_j^{(k)},\kappa))\Bigg\}\mathbf{1}\{y_j>c\}.
\end{aligned}
$$

We assume $F_M(0^+;\boldsymbol{\Psi}^{(k)})=0$ in the following derivations, which is generally the case.

4.2.2 M-step

The M-step on the $(k+1)$th iteration entails solving the following system of three equations:

$$
\begin{aligned}
&\sum_{j=1}^{n}\sum_{i=1}^{g}\frac{\partial}{\partial\mathcal{W}}\left[\tau_{ij}(y_j;\boldsymbol{\Psi}^{(k)},\boldsymbol{x}_j)\log(\pi_{ij})\right]\Big|_{\mathcal{W}=\mathcal{W}^{(k)}}\mathbf{1}\{0<y_j\le c\}=\mathbf{0}, \\
&\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}(y_j;\boldsymbol{\Psi}^{(k)},\boldsymbol{x}_j)\frac{\partial}{\partial\mathcal{B}}\left[\log f_i(y_j;\boldsymbol{\beta}_{ij},\phi_i,\boldsymbol{x}_j)-\log F_M(c;\boldsymbol{\Psi},\boldsymbol{x}_j)\right]\Big|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k)}}\mathbf{1}\{0<y_j\le c\}=\mathbf{0}, \\
&\sum_{j=1}^{n}\frac{\partial}{\partial\boldsymbol{\gamma}}\left[\log f_T(y_j;\boldsymbol{\gamma}_j,\kappa,\boldsymbol{x}_j)-\log(1-F_T(c;\boldsymbol{\gamma}_j,\kappa,\boldsymbol{x}_j))\right]\Big|_{\boldsymbol{\gamma}_j=\boldsymbol{\gamma}_j^{(k)}}\mathbf{1}\{y_j>c\}=\mathbf{0}.
\end{aligned}
\tag{4.13}
$$

Recall that π_{ij}'s are functions of $\mathcal{W}$, τ_{ij} is functions of π_{ij}. The first equation in (4.13) can be solved using a similar algorithm for logistic regression to produce updated estimates of $\mathcal{W}^{(k+1)}$ for the logistic regression coefficients as it represents the probabilities between 0 and 1. Concerning the computation of $\mathcal{B}$ and $\boldsymbol{\gamma}$ and applying the chain rule of McCullagh and Nelder (2019), the likelihood equation for $\boldsymbol{\gamma}$ given by the third equation of (4.13), conditional on $y_j>c$, can be expressed as

$$\sum_{j=1}^{n}w(\mu_j)(y_j-\mu_j)\eta_j'(\mu_j)\boldsymbol{x}_j=\mathbf{0}, \tag{4.14}$$

where $\mu_j = \exp(\boldsymbol{\gamma}^T \boldsymbol{x}_j)$, η_j is the log-link function with format given by (4.6) and $w(\mu_j)$ is the weight function defined by

$$w(\mu_j) = \frac{1}{[\eta'_j(\mu_j)]^2} V(\mu_j),$$

where $V(\mu_j)$ represents the variance function of μ_j presented in Section 4.1.2. It can be seen that for fixed κ, the likelihood equation for $\boldsymbol{\gamma}$ is independent of κ.

The equation (4.14) can be solved using iteratively reweighted least squares (IRLS) (Nelder and Wedderburn, 1972). The adjusted response variable $\tilde{y}_j$ for the $(k+1)$th iteration is given by

$$\tilde{y}_j^{(k)} = \eta\left(\mu_j^{(k)}\right) + \left(y_j - \mu_j^{(k)}\right)\eta'_j\left(\mu_j^{(k)}\right), \qquad j = 1, 2, \ldots, n. \tag{4.15}$$

These n adjusted responses are then regressed on the covariates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ using weights $w(\mu_1^{(k)}), w(\mu_2^{(k)}), \ldots, w(\mu_n^{(k)})$. Convergence can be obtained with a sequence of likelihood values that are bounded above as Dempster (1968) show that a local maximum can be found if (very weak) conditions that $Q(\boldsymbol{\Psi})$ is continues in $\boldsymbol{\Psi}$ and component densities of mixing proportions are specified. This produces an updated estimate $\boldsymbol{\gamma}^{(k+1)}$ for $\boldsymbol{\gamma}$ and, consequently, the updated estimates $\mu_j^{(k+1)}$ for the μ_j for use in the right-hand side of (4.15) to update the adjusted responses, and so on. This procedure is repeated until the variations in the estimates are small enough.

Same as (4.14), the likelihood for $\mathcal{B}$ given by the second equation in (4.13), conditional on $0 < y_j \leq c$, can be written as

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij}\left(y_j; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_j\right) w(\mu_{ij})(y_j - \mu_{ij})\eta'_i(\mu_{ij})\left[\frac{\partial}{\partial \mathcal{B}}\eta_i(\mu_{ij})\right] = \mathbf{0} \tag{4.16}$$

where μ_{ij} is the mean of Y_j for the ith component, η_i has similar structure as (4.6). Given that

$$\frac{\partial}{\partial \boldsymbol{\beta}_h}\eta_i(\mu_{ij}) = \begin{cases} \boldsymbol{x}_j, & \text{if } h = i \\ 0, & \text{otherwise} \end{cases},$$

equation (4.16) reduces to solving

$$\sum_{j=1}^{n} \tau_{ij}\left(y_j; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_j\right) w(\mu_{ij})(y_j - \mu_{ij})\eta'_i(\mu_{ij})\boldsymbol{x}_j = \mathbf{0}, \tag{4.17}$$

separately, for each $\boldsymbol{\beta}_i$ to produce $\boldsymbol{\beta}_i^{(k+1)}$, $i = 1, 2, \ldots, g$. Similar to (4.15), responses $y_1, y_2, \ldots, , y_n$ are fitted with weights $\tau_{i1}(y_1; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_1), \ldots, \tau_{in}(y_n; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_n)$ and fixed dispersion parameter ϕ_i. Equation (4.17) can then be solved using the IRLS approach for a single GLM. The double summation over i and j in (4.16) can be handled by expanding the response vector to have dimension $g \times n$ by replicating each original observation $(y_j; \boldsymbol{x}_j^T)^T$

by g times, with weights $\tau_{i1}(y_1; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_1), \ldots, \tau_{in}(y_n; \boldsymbol{\Psi}^{(k)}, \boldsymbol{x}_n)$, fixed dispersion parameters $\phi_1, \phi_2, \ldots, \phi_g$, and linear predictors $\boldsymbol{x}_j^T \boldsymbol{\beta}_1, \boldsymbol{x}_j^T \boldsymbol{\beta}_2, \ldots, \boldsymbol{x}_j^T \boldsymbol{\beta}_g$.

4.2.3 Specification of Parameters

The splicing probabilities p_1 and p_2 are estimated outside of E-M steps. Recall that $p_1(\boldsymbol{\alpha}, c)$ $p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})$ represents proportion of zero loss incidents with covariates $\boldsymbol{x_j}$, which is also a function of $\boldsymbol{\alpha}$ and c (see equation (4.7)). After setting a proper splicing point c, p_1 can be defined as a logistics regression model that describes the probability of zero loss incidents happening as a function of covariates $\boldsymbol{x_j}$ and coefficients $\boldsymbol{\alpha}$, given by

$$p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}) = \frac{1}{1 + e^{-\boldsymbol{x}_j^T \boldsymbol{\alpha}}}. \tag{4.18}$$

With the logistic model, estimates of p_1 are always between 0 and 1.

Let Y_j^* be a random variable, with $Y_j^* = 1$, if $Y_j = 0$, and $Y_j^* = 0$, if $Y_j > 0$, so Y_j^* is a Bernoulli random variable. More specifically, assume that

$$\Pr\{Y_j^* = y_j^* | \boldsymbol{x_j}\} = p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})^{y_j^*} (1 - p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}))^{1-y_j^*},$$

where $\boldsymbol{x_j}$ is the covariates of Y_j. Clearly, $\Pr\{Y_j^* = 1 | \boldsymbol{x_j}\} = p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})$, representing the probability that an observation with risk $\boldsymbol{x_j}$ has zero losses, where $p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})$ is given by (4.18). The likelihood function for all the n observations $\boldsymbol{y}$ can be expressed as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}; \boldsymbol{y}, \boldsymbol{x}) &= \prod_{j=1}^{n} p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})^{y_j^*} (1 - p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}))^{1-y_j^*} \\ &= \prod_{j=1}^{n} \left(\frac{e^{\boldsymbol{x}_j^T \boldsymbol{\alpha}}}{1 + e^{\boldsymbol{x}_j^T \boldsymbol{\alpha}}} \right)^{y_j^*} \left(\frac{1}{1 + e^{\boldsymbol{x}_j^T \boldsymbol{\alpha}}} \right)^{1-y_j^*} \end{aligned} \tag{4.19}$$

Since we have in total 36 risk level combinations for covariates and $\boldsymbol{x_j}$ must be one of them, the likelihood function (4.19) can be re-written as

$$\mathcal{L}(\boldsymbol{\alpha}; \boldsymbol{y}, \boldsymbol{x}) = \prod_{i=1}^{36} \left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\alpha}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\alpha}}} \right)^{n_{0,i}} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\alpha}}} \right)^{n_i - n_{0,i}}$$

where $n_{0,i}$ is the number of observations of zero losses with ith risk combination $\boldsymbol{x_i}$, and n_i is the number of observations with ith risk combination such that $\sum_{i=1}^{36} n_i = n$.

The corresponding log likelihood function can be obtained as

$$\ell(\boldsymbol{\alpha}; \boldsymbol{y}, \boldsymbol{x}) = \sum_{j=1}^{36} \left[n_{0,i} \cdot \boldsymbol{x}_i^T \boldsymbol{\alpha} - n_i \log \left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\alpha}} \right) \right].$$

The maximum likelihood estimation of $\boldsymbol{\alpha}$ can be obtained by maximizing the above log likelihood, which has no closed-form solutions. Therefore, a technique like the iteratively reweighed least squares can be used to find an estimate of the regression coefficients (O'Leary, 1990).

In order to have the overall distribution to be continuous at the splicing point, we set at c such that, for the density function,

$$\lim_{y \to c_-} f_Y(y; \boldsymbol{\alpha}, \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \kappa, \boldsymbol{x}) = \lim_{y \to c_+} f_Y(y; \boldsymbol{\alpha}, \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \kappa, \boldsymbol{x})$$

Using (4.7), we get

$$p_2(\boldsymbol{\alpha}, \boldsymbol{x_j}) \frac{f_M(c; \boldsymbol{x}, \mathcal{B}, \boldsymbol{\Phi})}{F_M(c; \boldsymbol{x}, \mathcal{B}, \boldsymbol{\Phi})} = [1 - p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}) - p_2(\boldsymbol{\alpha}, \boldsymbol{x_j})] \frac{f_T(c; \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\theta})}{1 - F_T(c; \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\theta})}, \tag{4.20}$$

and by (4.20), we can then determine $p_2(\boldsymbol{\alpha}, \boldsymbol{x_j})$ which is given by

$$p_2(\boldsymbol{\alpha}, \boldsymbol{x_j}) = \frac{[1 - p_1(\boldsymbol{\alpha}, \boldsymbol{x_j})] \frac{f_T(c; \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\theta})}{1 - F_T(c; \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\theta})}}{\frac{f_M(c; \boldsymbol{x}, \mathcal{B}, \boldsymbol{\Phi})}{F_M(c; \boldsymbol{x}, \mathcal{B}, \boldsymbol{\Phi})} + \frac{f_T(c; \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\theta})}{1 - F_T(c; \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\theta})}}.$$

Dispersion parameters $\boldsymbol{\phi}$ and κ for f_M and f_T, respectively, in (4.7) are pre-determined before E-M steps by the method of moments trained on body part and tail part exclusively. They are normally estimated by the programming packages.

Until now, we have considered the fitting of a Zi-MCR model with a given value of severity threshold c value. Typically, where our model is being used to handle overdispersion, the value of c can be predetermined from data using extreme value analysis or expert opinion via performing cyber insurance policy limit and similar matters. This is primarily motivated by estimation stability, which is adopted by Reynkens et al. (2017). Furthermore, conducting formal tests at any stage of this sequential process is challenging because regularity conditions for the likelihood ratio test statistic's typical asymptotic null distribution do not hold. As the constrained likelihood function needs to be optimized with respect to the nuisance parameters of the model, even for the small dimension of the space of parameters of interest in simple models, the computational burden can be extensive (Peel and MacLahlan, 2000). Observing the trend in the log-likelihood as c is increased from a sequence of severity levels $1000, 5000, 10000, 50000$ and 100000 can provide us with a heuristic for determining the optimal value of c. When dealing with a data-driven model, this method for selecting a splicing point makes more sense and is widely used (Gan and Valdez, 2018).

4.3 Analysis of Cyber Breach Chronology Data

In this Section, we illustrate the efficiency of the EM algorithm on estimation by fitting a Zi-MCR model, as proposed in Section 4.1.3, to the PRC dataset. Furthermore, modeling specification and covariates are discussed, and several distribution combinations are tested to select one with the best performance.

This study is based upon the PRC cyber breach incident data by stratifying the residuals. The training set fine-tunes all candidate models, and their performance and out-of-sample validation are checked upon the test set. We conduct 5-fold cross-validation and set 80% as the training data to fit the models. Based on a set of breach observations, the problem is to estimate whether the unknown parameters can be obtained in the vector $\boldsymbol{\Psi} = (\boldsymbol{\alpha}^T, \mathcal{W}^T, \mathcal{B}^T, \boldsymbol{\gamma}^T)^T$, as in (4.10). We represent the logarithm rescaled number of loss records of data breached due to cyber incidents explained in Section 2.3.1 as a target or dependent random variable Y_j, and nine covariates, including intercept, two business levels, two breach levels, and four location area levels, as described in the Sections 2.3.2 and 2.3.3 as vector coefficients $\boldsymbol{x}_j$, for $j = 1, 2, \ldots, 9$.

Table 4.1 displays the summaries of three categorical variables; the proportion of zeros and differences between the mean values of the categories numerically illustrate their distribution patterns. These results demonstrate the importance of letting splicing weights depend on covariates and separately modelling body and tail parts.

Feature	Category	Proportion of Zeros	Non-Zero Mean	Total Count
	Medical	652 (15.7%)	58501	4161
Organizations	Businesses	1434 (63.0%)	2197387	2275
	Non-businesses	577 (34.8%)	174932	1659
	External Malicious	1125 (44.1%)	1635354	2549
Breaches	Internal Malicious	775 (31.8%)	462591	2440
	Internal Negligent	763 (24.6%)	75805	3106
	Area 1	1064 (35.2%)	390949	3024
	Area 2	143 (26.9%)	301578	531
Territories	Area 3	449 (27.3%)	666947	1642
	Area 4	283 (25.9%)	389733	1093
	Area 5	724 (40.1%)	1478791	1805

Table 4.1: Summary of categorical variables

In the PRC dataset, a source of the heterogeneity is mainly from businesses that have or do not have high prevention defence systems and active cyber risk managing activities, such as healthcare and financial service organizations. This is explained in Section 2.3.2 by comparing their kernel plots and enterprise features. The body part component may be viewed as two groups corresponding to whether those incidents happened within medical organizations. The problem is to estimate the medical and non-medical organization mixture rate, that is, the mixing proportion π_1. Given $g = 2$, $\boldsymbol{\alpha}^{(k+1)}$ and $\mathcal{W}^{(k+1)}$ can be calculated using binomial error structure with the canonical logit transformation as the link. For illustrative purposes, we fit several popular distribution combinations on a mixture of body and heavy tail parts. To measure the overall goodness of fit of those fitted distributions, we calculate the Akaike information criterion (AIC) statistics. Table 4.2 reports the global fit distributions overall AIC values on a given $c = 5000$ threshold. The fit from Lognormal and Weibull body mixture and Pareto tail outperforms with the lowest AIC. We conduct a simulation study based on the entire data set to comprehend further the adaptability of the model chosen. The procedure is repeated 200 times to ensure a thorough analysis of the chosen distribution combination. The estimated parameters are summarized in Table 4.3 using the distribution combined with the lowest AIC value, Lognormal-Weibull for the body part and Pareto for the tail part.

Body		Tail	AIC	Body		Tail	AIC
Gamma	Lognormal	Pareto	-46.3333	Gamma	Lognormal	Lognormal	-51.5420
Gamma	Gamma	Pareto	-55.8390	Gamma	Gamma	Lognormal	-49.0340
Lognormal	Weibull	Pareto	$\mathbf{-56.6044}$	Lognormal	Weibull	Lognormal	-45.6592
Gamma	Weibull	Pareto	-55.8390	Gamma	Weibull	Lognormal	-47.3896

Table 4.2: Overall goodness-fit

As an example, we express below the explicit density function of our Zi-MCR model with the lowest AIC; it is given by

$$\begin{aligned} & f_Y(y_j; \boldsymbol{\alpha}, \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \kappa, \boldsymbol{x}_j) \\ & = p_1(\boldsymbol{\alpha}, \boldsymbol{x}_j)\mathbf{1}\{y_j = 0\} \\ & \quad + p_2(\boldsymbol{\alpha}, \boldsymbol{x}_j)\frac{\sum_{i=1}^{2} \pi_{ij}(\mathcal{W}, \boldsymbol{x_j}) f_i\left(y_j; \exp\left(\boldsymbol{\beta}_i^T \boldsymbol{x}_j\right), \phi_i\right)}{F_M(c; \mathcal{W}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi_1, \phi_2, \boldsymbol{x_j})}\mathbf{1}\{0 < y_j \leq c\} \\ & \quad + [1 - p_1(\boldsymbol{\alpha}, \boldsymbol{x_j}) - p_2(\boldsymbol{\alpha}, \boldsymbol{x_j})]\frac{f_T\left(y_j; \exp\left(\boldsymbol{\gamma}^T \boldsymbol{x}_j\right), \kappa\right)}{1 - F_T(c; \boldsymbol{\gamma}, \kappa, \boldsymbol{x_j})}\mathbf{1}\{y_j > c\}, \end{aligned} \tag{4.21}$$

where f_1 is the Lognormal density function with scale (dispersion) parameter ϕ_1 and coefficients $\boldsymbol{\beta}_1$, f_2 is Weibull density function with scale parameter ϕ_2 and coefficients $\boldsymbol{\beta}_2$, f_T is type II Pareto density function with scale parameter κ and coefficients $\boldsymbol{\gamma}$, and F_M is cumulative distribution function of the mixture of Lognormal and Weibull distributions, with the following forms, respectively:

$$\begin{aligned} f_1(y_j; \boldsymbol{\beta}_1, \phi_1) &= \frac{\phi_1}{y_j\sqrt{2\pi}} \exp\left(-\frac{\phi_1^2\left[\ln(y_j) - \exp\left(\boldsymbol{\beta}_1^T \boldsymbol{x}_j\right)\right]^2}{2}\right), \qquad y_j > 0 \\ f_2(y_j; \boldsymbol{\beta}_2, \phi_2) &= \exp\left(\boldsymbol{\beta}_2^T \boldsymbol{x}_j\right)\phi_2\left(y_j\phi_2\right)^{\exp\left(\boldsymbol{\beta}_2^T \boldsymbol{x}_j\right)-1} \exp\left[-\left(y_j\phi_2\right)^{\exp\left(\boldsymbol{\beta}_2^T \boldsymbol{x}_j\right)}\right], \qquad y_j > 0 \\ f_T(y_j; \boldsymbol{\gamma}, \kappa) &= \exp(\boldsymbol{\gamma}^T \boldsymbol{x}_j)\kappa\left(1 + y_j\kappa\right)^{-[\exp(\boldsymbol{\gamma}^T \boldsymbol{x}_j)+1]}, \qquad y_j > 0, \end{aligned}$$

Vector	Coefficients	Estimation	Vector	Coefficients	Estimation
$\boldsymbol{\alpha}^T$	α_1	-1.2292	$\boldsymbol{\gamma}^T$	γ_1	5.2892
	α_2	-0.3876		γ_2	0.6577
	α_3	-0.0915		γ_3	-0.1476
	α_4	0.5934		γ_4	-0.6063
	α_5	1.6082		γ_5	1.1548
	α_6	-0.1650		γ_6	-0.4896
	α_7	-0.2233		γ_7	0.7618
	α_8	-0.1999		γ_8	1.4402
	α_9	-0.6482		γ_9	1.5340
$\boldsymbol{\beta}_1^T$	β_{11}	1.1170	$\boldsymbol{\beta}_2^T$	β_{21}	2.1593
	β_{12}	0.9828		β_{22}	0.1305
	β_{13}	-0.3131		β_{23}	-0.2645
	β_{14}	0.0360		β_{24}	0.0192
	β_{15}	0.0404		β_{25}	-0.2291
	β_{16}	-0.4804		β_{26}	0.1814
	β_{17}	0.4161		β_{27}	0.2837
	β_{18}	0.2168		β_{28}	0.2892
	β_{19}	0.0734		β_{29}	0.2688
$\mathcal{W}^T$	w_1	0.1704	$\boldsymbol{\phi}$	ϕ_1	1.1112
	w_2	0.9999		ϕ_2	0.0167
	w_3	0.5000	κ	κ	1.1765
	w_4	0.2856			
	w_5	0.1557			
	w_6	0.6757			
	w_7	1.0000			
	w_8	1.0000			
	w_9	0.0027			

Table 4.3: Parameter estimations

and

$$
\begin{aligned}
&F_M\left(c; \mathcal{W}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi_1, \phi_2, \boldsymbol{x_j}\right) \\
&= \pi_{1j}\left(\mathcal{W}, \boldsymbol{x_j}\right) \Phi\left(\phi_1\left[\ln(c) - \exp(\boldsymbol{\beta}_1^T \boldsymbol{x}_j)\right]\right) + \pi_{2j}\left(\mathcal{W}, \boldsymbol{x_j}\right)\left(1 - \exp(-c\phi_2)^{\exp\left(\beta_2^T \boldsymbol{x}_j\right)}\right).
\end{aligned}
$$

We take three representative risk combinations as illustrative examples and plot their PDF upon logarithm scale of positive loss severity as shown in Figures 4.1, 4.2 and 4.3. The limits of Y-axis is scaled down according to the range of density probabilities and vertical white dots represents the logarithm of splicing point/threshold c, log(5000), with the point mass at zero (proportion of zeros) excluded in these figures. The vertical white dotted lines acts as the divider of splicing point with number represents cumulative percentage of mixture component part. Severe loss of business external malicious breach activities in South area leads to a heavy right tailed density function and overall huge mean showed in Figure 4.1. Figure 4.2 shows the density function of non-business internal malicious in Midwest that has a moderate right tail representing 25% of positive losses. While the density of medical internal negligent in Northeast, showed in Figure 4.3, performs differently with majority of low-severity incidents which represent about 91% positive losses, 62.7%[3] in terms of cumulative loss, and thus the density function has a very thin tail. Those results and findings would be echoed on some level in Chapter 5 loss aggregation analysis.

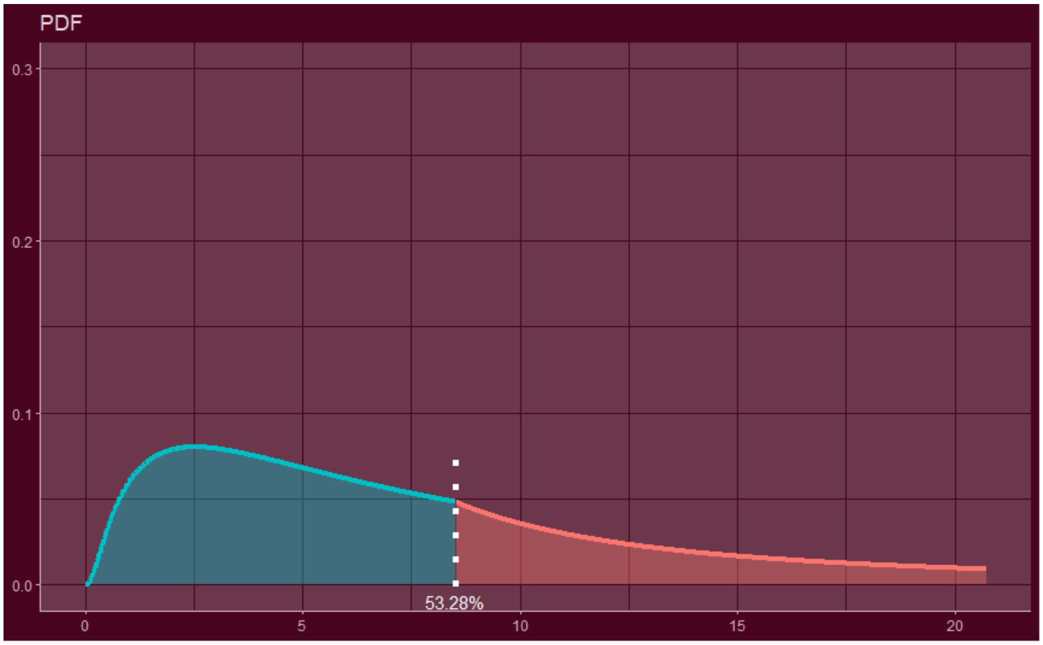


Figure 4.1: Business-External Malicious-South

[3] 62.7% = (1 − 31.3%) × 91.35%, where 31.3% is proportion of zero of this combination.

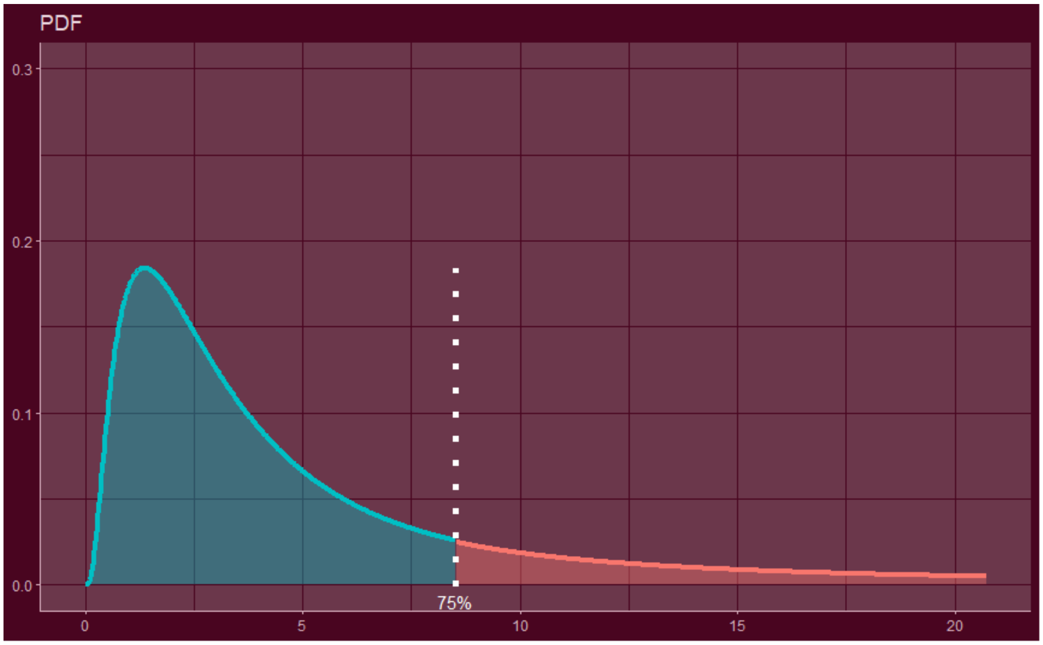


Figure 4.2: Nonbusiness-Internal Malicious-Midwest

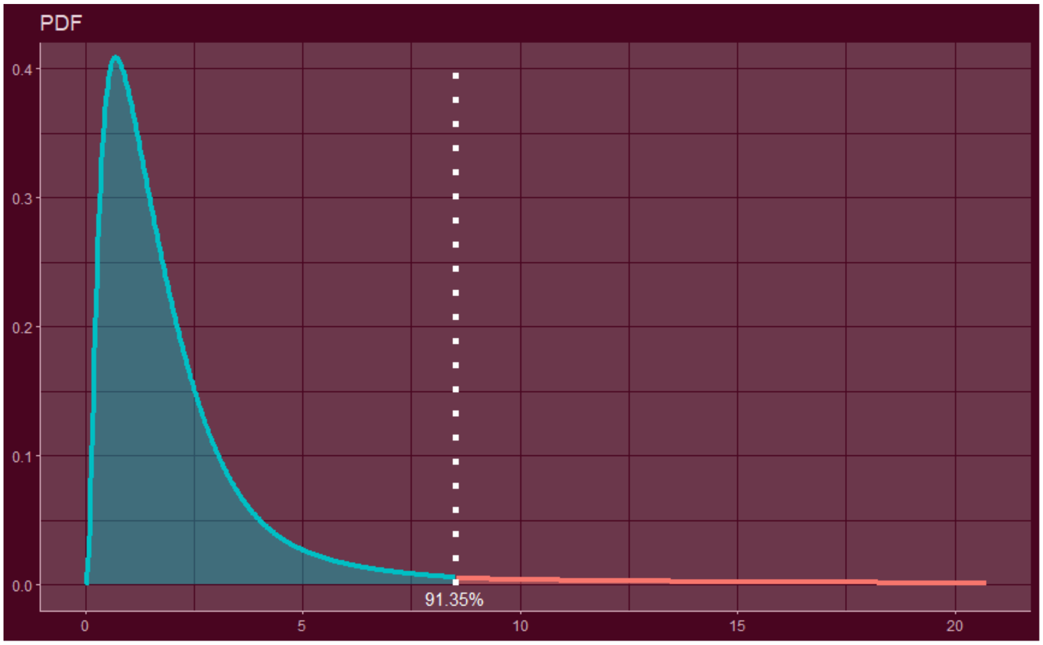


Figure 4.3: Medical-Internal Negligent-Northeast

Since the metric is based on data-driven analysis, we can draw a conclusion that the selected combination, Lognormal-Weibull and Pareto, has the most explanation power of PRC dataset. If other dataset is given, methodology and algorithm in generating the metric should not change, selected distribution combination would depend on a case by case basis. In Chapter 5, We illustrate an application of our proposed severity methodology in examining aggregate cyber losses by combining the frequency modeling approach proposed in Chapter 3 based on the same dataset.

4.4 Discussion

Once an insurance loss model has been constructed, in addition to applying techniques to data sets, we must consider numerous modelling-related factors, such as risk management and pricing decisions (for insurers) or the impact on capital requirements (for enterprises). Our model and findings provide meaningful insights to risk mitigation and risk transfer techniques, which benefits not only the individual organization, but also the overall economy.

Cyber risk exists because computer data is valuable to individuals, business, and governments; therefore, the data must be protected by organizations that store privileged information. Financial firms receive, maintain, and store large amounts of personally identifiable information. Recent security research (Varonis, 2021) indicates that most businesses have unprotected data and inadequate cybersecurity practices, making them susceptible to data loss. As more executives and decision-makers recognize the value and significance of security investments, cybersecurity budgeting has steadily risen to successfully combat potential digital property loss. A systemic cyber event could cost multiple times the current risk retention estimate. As a result of regulatory scrutiny and the need for improved portfolio management, businesses conduct scenario modelling and sensitivity tests regularly based on their changing risk appetite. To reduce cyber risk, organization can adopt threshold limits by monitoring risk with preset limit based on established risk criteria, trigger will be placed in threshold has been breached. The objective is to achieve and maintain an acceptable level of risk at a reasonable cost. Under the leadership of the Chief Risk Officer, companies must revise their strategies, including changes to their risk appetite and the composition of their hedge products. Due to some businesses' nature or responsibilities, increasing risk appetite or security investments may not be sufficient to achieve the risk management objective. Such limited reserved retention can have disastrous financial consequences if a data breach occurs, forcing the organization to absorb the costs associated with internal remediation and its liability to third parties. In this perspective, cyber insurance has become an effective alternative or backup tool for managing cyber risk.

Our investigation of large claims and an excess of zeros raises the issue of the insurability of cyber risks under various feature characteristics. To eliminate the variance caused by the heavy tail, the non-catastrophic loss can then be used to train a predictive model.

Reinsurance will kick in if the loss exceeds the company's tolerance level to ensure that insurers are not severely impacted. Our model offers an additional perspective on coping with extreme loss values, such as those caused by cyber attacks. Risk selection is one of the most crucial processes when designing an insurance product. Since not all customers are equally attracted to an insurance product, segmenting the risks into distinct groups is advantageous to prevent adverse selection. To ensure that cyber insurance products are priced appropriately, we use the results from the Section 4.3 to divide risks into categorizable segments. In addition, our model can be utilized to perform a preliminary pre-screening of a prospective client to facilitate rate discrimination and the creation of customized contracts. This security audit enables the insurer to capitalize on the profit opportunity presented by the interdependence of cyber risks.

Combining the loss frequency and severity distribution through convolution is a conventional method for estimating the aggregate loss distribution. Given the proposed mixture and composite severity model, aggregate losses can be estimated through simulations since our proposed frequency model is semi-parametric with a simulated posterior distribution without a closed form of distribution. In Chapter 5, We will show the estimation of the aggregation of cyber losses and discuss the insurance applications.

The financial sector faces cybersecurity risks in their daily operations while insuring product providers. Insurers receive privacy information such as personal health and financial situation from both policyholders and claimants. The cost of cyber insurance increased by an average of 96% in the third quarter of 2021 as organizations faced a daily onslaught of cyberattacks (McLennan, 2021). To mitigate the premium price increase, policyholders may increase their retention level. As a result, insurers must improve predictive analysis and cyber risk models to maintain market share and company solvency. Our model provides a method for measuring cyber risk severity, and there are multiple ways to extend this method. As previously stated, all of our results are based on the assumption of equal exposure, whereas exposure is the most crucial factor in determining the pure premium. Cyber risk loss exposures are any conditions that present the possibility of financial loss to an organization from property, net income and liability as a consequence of advanced technology transmissions, operations, maintenance, development and support. Training the predictive model under the assumption of equal exposure in a defined time period would be an important direction for future research once prior experience data with exposure information is obtained.

Chapter 5

Cyber Loss Aggregation and Industry Application

Based on frequency and severity modeling structures studied in Chapters 3 and 4, we employ their results for our data-driven analysis and propose several approaches in generating aggregate loss and implementation strategies that can be utilized by insurance industry. Starting with presenting general form of cyber loss aggregation model and introducing risk measures in Section 5.1, we introduce simple loss aggregation approach assuming that the loss frequency and loss severity are independent and severity is not random in Section 5.2, and Markov chain Monte Carlo (MCMC) loss aggregation approach where the loss frequency is dependent on the average loss severity and the loss severity has its own selected zero-inflated mixture and composite regression model in Section 5.3. The impact of applying different deductibles, limits and reinsurance practice are discussed in Section 5.4. Finally, applications of compound loss aggregation to current U.S. cyber insurance market are discussed in Section 5.5.

5.1 Cyber Loss Aggregation Model and Risk Measures

The premium calculation algorithm, known as rate order calculation, is applied to categorize segmentation to derive final premium rates. In order to set competitive premiums and develop sustainable underwriting plans, insurers extensively use historical loss data to seek economies of scale and premium balancing. Statistical algorithms and mathematical modelling arguments are used to structure aggregate cyber risks. The purpose of this section is to describe an aggregate loss model based on the total amount of cyber loss that occurs in a quarter concerning a group of different risk characteristics and apply this model to determining increased limit factors (ILFs) based on the underlying data in order to balance statistical and economic constraints. According to risk theory (Bühlmann, 2007), a collective risk model with aggregate loss S, which represents the total amount of quarterly loss due to cyber attacks in our study, can be defined as

$$S = \begin{cases} \sum\limits_{k=1}^{N} Y_i & N > 0, \\ 0 & N = 0, \end{cases}$$

where loss counts (cyber attacks) N and non-negative loss amounts (records/data breached), $Y_1, Y_2, \ldots, Y_N$, are random variables and are assumed independent; that is, N does not rely on the loss severity. Note that the individual loss amount can be zero with certain probability in our case (breach happens but no incurred loss) and or be positive, which are assumed to be independent and identically distributed following a unified distribution including the point mass at zero.

For this collective risk model, the expected value and variance of the aggregate claims S are as follows:

$$E(S) = E(N)E(Y), \tag{5.1}$$

$$\mathrm{Var}(S) = E(N)\mathrm{Var}(Y) + \mathrm{Var}(N)[E(Y)]^2.$$

Due to the complexity of our proposed statistical model/distribution for the total loss, in practical applications the following premium calculation principles are typically used as approximations or references for the determination of the premiums. In our study, where loading factors are not specified, pure premium creates an example, which can be modified once loading information is obtained from businesses. Here we list some well-known moment based premium calculation principles; all these are based on the mean only or both the mean and the variance of the loss counting random variable N and the loss severity random variable Y, which can be determined relatively easily.

- Pure Risk Premium

$$P = E(N)E(Y); \tag{5.2}$$

- Premium with Safety Loading Factor θ

$$P_{SL}(\theta) = (1+\theta)E(N)E(Y), \qquad \theta \geq 0;$$

- Premium with Variance Loading Factor a

$$P_V(a) = E(N)E(Y) + a[E(N)\text{Var}(Y) + E(Y)^2\text{Var}(N)], \qquad a \geq 0;$$

- Premium with Standard Deviation Loading Factor b

$$P_{SD}(b) = E(N)E(Y) + b\sqrt{E(N)\text{Var}(Y) + E(Y)^2\text{Var}(N)}, \qquad b \geq 0.$$

More premium principles are described in Bühlmann (1980), which are applied to a loss distribution to determine an appropriate premium to charge for the risk. The Risk Loading factors represent the percentage of insurance premium deducted from the premium payments to cover policy expenses and the variability of the loss, which act as a cushion against adverse experience.

In order to utilize the characteristics of the loss distribution for pricing, reserving and risk management, a well established loss distribution, either parametrically, non-parametrically, analytically, or by Monte Carlo simulation, is necessary. A risk measure is a functional mapping of loss distribution to real numbers encapsulating the risk associated with that loss distribution, which is an important tool in actuarial process.

The **Value at Risk (VaR)** risk measure in actuarial quantile premium principle is the loss at a certain probability level, which is specified with a given level α, typically 95% or 99%, denoted by VaR_α. More specifically, VaR_α represents the loss with probability α will not be exceeded (Rockafellar and Uryasev, 2002); mathematically, it is defined by

$$\Pr[L \leq \text{VaR}_\alpha] = \alpha.$$

That is, $\text{VaR}_\alpha = F_L^{-1}(\alpha)$, where $F_L(x)$ is the cumulative distribution function of the loss random variable L. Several insurance risks exhibit a heavier tail than the normal distribution, and VaR captures these potential losses, which offers a closer approximation of risk profile.

Since the quantile risk measure does not consider what the loss will be if $1-\alpha$ worst case event occurs, it fails to reflect the loss distribution above the quantile. The **Conditional Tail Expectation (CTE)**, also known as Tail Value at Risk (TVaR) or Expected shortfall (ES), was proposed to address this problem (Hardy, 2006). Like the VaR, the CTE is defined using confidence level α as well, denoted by CTE_α, which represents the expected loss given that the loss falls in the worst $(1-\alpha)$ part of the loss distribution. Mathematically, it can be defined, given the α-quantile risk measure VaR_α as

$$\mathrm{CTE}_\alpha = E[L|L > \mathrm{VaR}_\alpha].$$

CTE provides a more precise representation of exceptional events that could pose a threat to the financial standing of the firm.

Estimates of risk measures such as VaR and CTE can be analyzed using the aggregate loss modeling results at a given risk tolerance level. There are two commonly used approaches to generate risk measures and related quantities. One is the variance-covariance/parametric approach, where variance and covariance are estimated using historic data and the target distribution is transformed to a multivariate normal distribution. It is easy to implement once L is a closed form with pre-defined assumptions. If the case requires more flexible with subjective judgement and information come into play, Monte Carlo approach seems to be a proper choice. It uses a simulation process to generate large enough possible outcomes as long as the probability distribution of risk factors and their co-movements are defined. This procedure is to be illustrated in Section 5.3.

5.2 Loss Aggregation Under a Simplified Model Setting

The purpose of the case study in this section is to examine the effect of insurance coverage modifications such as deductibles, policy limits and their combinations in insurance application under a simplified aggregate model. In actuarial literature, a well-known and particularly practical method, called the frequency-severity method (Friedland, 2010), estimates the insurance costs based on the expected number of claims determined by the claim frequency model and the average claim cost from the severity model, both developed and fitted using the relevant historical data. With the same idea, for this case study we use the frequency model proposed in Chapter 3 and use the empirical average of the severity with particular risk factors instead of the expected cost of a claim from the model for loss amounts. Under this simplified model setting, we can nevertheless obtain a distribution for the quarterly total loss for a specified time period. This approach has several advantages: changes over time can be monitored and attributed to frequency or severity. For example, we have showed that cyber loss frequency has a significant polynomial seasonal pattern. As data accumulated by time, our methodology can be quickly updated to fit model components, which can be done by training the model again based on our methodologies.

Our ultimate goal is to estimate the monetary loss caused by the cyber attack incidents and the associated data breaches. In this case study and the severity model proposed in Chapter 5, the severity/loss refers to the number of data breaches recorded. We then convert units of loss recorded into their corresponding monetary loss amount using the following relationship/rule developed in Jacobs (2014):

$$\ln(\text{dollar amount loss}) = 7.68 + 0.76 \times \ln(\text{loss records breached}), \tag{5.3}$$

where ‘loss records breached’ refers to the number of records impacted. This relationship has been used in (Eling and Loperfido, 2017) to estimate prices for cyber insurance policies and to provide useful insights for actuarial applications. Romanosky (2016) develops a more comprehensive model for cyber incidents that helps better understand the relevant factors driving costs based on a rich set of data with revenue, lawsuits and breach occurrence information. Since the dataset we use for this thesis does not contain such detailed components, we adopt the relatively simple relationship (5.3) for approximating the monetary loss for insurance applications.

Align with Chapter 3 and Chapter 4 rationales, the collective risk model for $(J+1)$th quarter aggregate loss S_{J+1} under the simplified model setting described above can be expressed as

$$S_{J+1} = \sum_{i=1}^{36} \left(N_{i,J+1} \cdot \bar{Y_i} \right), \tag{5.4}$$

where $N_{i,J+1}$ is the loss counts (cyber attack incidents) of the ith combination and $(J+1)$th quarter, and $\bar{Y_i}$ is the average of last three quarterly loss amounts for the ith combination for $i = 1, ..., 36$. These 36 combinations are formed based on featured covariate risk levels of the dataset studied including 3 business types, 3 breach types, and 4 location areas; see Table 5.2 for detailed descriptions. Aggregated quarterly loss can be obtained by adding all 36 combinations’ quarterly losses. Our previous loss frequency study reveals there exists seasonal pattern; consequently a polynomial covariate is included in regression model. The quarterly loss counts for the $(J+1)$th quarter with a specific feature combination can be predicted by setting polynomial covariate equal to $J+1$.

Using the posterior frequency distributions on characteristic segments obtained in Section 3.4 based on observations up to the Jth quarter, we can generate a set of total 36 aggregate loss distributions for all the level combinations. By using the frequency-severity technique described above, the aggregated quarterly loss distribution for the $(J+1)$th quarter, S_{J+1}, given by (5.4), can be obtained. We then apply log-log model (5.3) to convert the number of records breached into its corresponding dollar amount loss.

We illustrate our numerical results by considering only two representative geographical locations (northeast and west) and two business types (non-business and business). For each combination of risk factors, we first use the mean values of estimated coefficients of covariates based on their posterior distributions in NB-GLMM to generate corresponding frequency distribution. Then the aggregate loss distribution is obtained by treating the severity of loss as constant which equals to the latest three quarters average loss amounts (number of data breaches recorded). In this case, the variability of the aggregate loss is mainly contributed by frequency variation across different character combinations. In our case study, while loading factors are not specified, pure premium creates an example, which can be modified once loading information is obtained from businesses. Table 5.1 displays the total quarterly dollar amount loss caused by cyber attacks. Estimated loss amount is calculated according to (5.3) conversion method, in which the "loss records breached" amount is estimated by the following equation:

$$\text{Aggregated loss records breached} = E(l < S < u),$$

where l represents "Lower Threshold" and u represents "Upper Threshold" indicated in the Table, which can be utilized as deductible and policy limit, respectively, in terms of premium calculation.

Table 5.1: Quarterly aggregate loss in dollar amount.

Location	Business Type	Lower Threshold	Upper Threshold	Estimated Loss
Northeast	Business	-	-	197,891
		10,000	-	188,469
		-	10,000,000	197,891
		10,000	10,000,000	188,469
	Non-Business	-	-	2,283,023
		10,000	-	2,273,881
		-	10,000,000	1,164,335
		10,000	10,000,000	1,162,902
West	Business	-	-	1,408,541
		10,000	-	1,398,568
		-	10,000,000	1,264,013
		10,000	10,000,000	1,260,245
	Non-Business	-	-	14,661,661
		10,000	-	14,651,699
		-	10,000,000	1,680,241
		10,000	10,000,000	1,680,149

Based on these results, we have the following informative findings from different perspectives.

- There is a significant difference in dollar loss amounts between the Northeast and West regions, with about seven times larger in both Business and Non-business entities.
- For each respective region, non-business organizations face much higher cyber risks than business organizations do according to their more than ten times estimated loss differences in dollar amounts without coverage modification.

- Whether having a lower threshold makes no big difference in aggregate losses because nearly a same estimated loss amount with and without the $\$10,000$ threshold is observed.
- On the contrary, setting a higher threshold can reduce covered cyber losses gigantically in non-business organizations compared with that in business organizations.

Those insights are worth considering while setting premium rates and designing insurance products in order to reach an equilibrium covering limited risk by sufficient amount of premiums. A more comprehensive illustration based on our proposed frequency and severity models with discussions is provided in Section 5.3. More discussions regarding implementations of insurance coverage modifications and reinsurance with discussions on different perspectives can be found in Section 5.4.

5.3 Markov chain Monte Carlo (MCMC) Loss Aggregation

Monte Carlo simulation has been proved to be one of the most efficient approaches to determine the compound loss distribution in aggregating losses when the distributions for frequency and severity are not in a closed form (Cruz et al., 2015). However, a major challenge of applying this approach, especially in the Bayesian analysis, is the independent assumption of the loss frequency and the loss severity. In Section 5.2, we assumed that the frequency and the severity random variables are not conditional on each other, which is a simplified case illustration. According to the conclusion of Markov (1906), "Independence of quantities does not constitute a necessary condition for the existence of the law of large numbers". The Law of Large Numbers is a statistical concept that calculates the average number of events or risks in a sample or population to predict quantities of interest (Ewold, 1991). In order to tackle this limitation, MCMC serves a practical choice in generating aggregate loss with frequency dependent on severity, as iteration provides the possibility of simulating sequential samples via chain method. In Chapter 3, we propose a Bayesian negative binomial generalized linear mixed model (NB-GLMM) for the quarterly cyber incidents (frequency), where the average number of data breached (severity) over the past several quarters is used as one of the regressors. When generating samplers for the aggregate loss, the sequence of the average severity and frequency can be handled in a chain-dependent process (Katz, 1977). In this manner, a Markov chain Monte Carlo algorithm can be formalized to estimate/predict the aggregate loss of our interest.

It is natural to take a Bayesian approach, using our proposed Bayesian NB-GLMM model for the frequency and combining the zero-inflated mixture and composite regression model for the severity to simulate and analyze the aggregate loss within a given time period. The construction of the loss aggregation contains two phases. First phase is to simulate the loss frequency. The MCMC algorithm is applied in order to obtain a sample from the posterior distribution of frequency parameters. This can be carried out by cycling repeatedly through draws of each parameter conditional on the remaining parameters. To accomplish this, we need to sample from the conditional posterior distribution of each parameter via Gibbs sampler as described in Section 3.3.1. The stationary distribution of each parameter coming from this Markov chain finishes this phase by providing the joint posterior frequency distribution of interest. The second phase focuses on the generating loss amounts based on the corresponding loss counts simulated. With the same covariates classification and grouping, parameters are estimated using EM algorithm by fitting a Zi-MCR model as illustrated in Section 4.2. The loss severity is estimated by the Lognormal-Weibull and Pareto mixture components model which is approved to be the best fit from data-driven analysis in Section 4.3. Given a number of loss incidents simulated in phase one, loss severity amounts can be simulated correspondingly from this distribution. Aggregations of loss amounts thus constitute a quarterly compound loss distribution. We describe this approach with details below.

Let $\mathbf{D}_J$ denote the set of observations up to the Jth quarter, containing the number of cyber attack incidents and the number of data breaches (severity) incurred for all I combinations. Now, given $\mathbf{D}_J$, we are interested in predicting the distribution of $S_{i,J+1}$, for $i = 1, 2, \ldots, I$, where $S_{i,J+1}$ is the aggregate data breaches of the ith combination for the $(J+1)$th quarter and it can be expressed as

$$S_{i,J+1} = \sum_{l=1}^{N_{i,J+1}} Y_{i,l,J+1}\,, \tag{5.5}$$

where $N_{i,J+1}$ denotes the number of data breach incidents (frequency) of ith combination happened in $(J+1)$th quarter, and $Y_{i,l,J+1}$ represents its lth loss amount (severity), which is the number of records breached in the context of cyber data breaches.

The NB-GLMM model for $N_{i,j}$ under the Bayesian framework is presented in Chapter 3. For the purpose of this section, we recall the full model for $N_{i,j}$ and its Bayesian estimation of the parameters using Gibbs sampler and M-H algorithm. The full model can be described as follows:

$$\begin{aligned} N_{i,j}|\mathbf{x}, \mathbf{z}, \mu_{i,j}, \xi_j &\sim \mathcal{NB}(\mu_{i,j}, \xi_j) \\ \log(\mu_{i,j}) &= \boldsymbol{x}_{i,j}^T \boldsymbol{\beta}_j + \boldsymbol{z}_j^T \boldsymbol{b} \\ \boldsymbol{\beta}_j &\sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \\ \boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) \\ \boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}(\nu_0, \boldsymbol{S}_0^{-1}), \end{aligned}$$

in which the heterogeneity among the regression coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J$ is described by a multivariate normal distribution with mean $\boldsymbol{\theta}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$.

In Chapter 4, a zero-inflated mixture and composite regression model is proposed for the number of data breached (severity); its density function $f_Y(y; \boldsymbol{\alpha}, \mathcal{W}, \mathcal{B}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \kappa, \boldsymbol{x_i})$ is given by (4.7), in which the model parameters are estimated by EM algorithm based on observed data $\mathbf{D}_J$.

Utilizing a similar approach presented in Baumgartner and Jones (2015), we propose the following steps to simulate total loss for all the combinations in the $(J+1)$th quarter, $S_{i,J+1}$, $i = 1, 2, \ldots, I$, given by (5.5), where the parameters for $N_{i,j}$ are from their posterior distributions based on past data $\mathbf{D}_J$.

For each of the last K MCMC iterations (we take $K = 200$), $k = 1, 2, \ldots, 200$, and for each of the I combinations, $i = 1, 2, \ldots, I$, perform the following steps.

1. Generate $\boldsymbol{\beta}_{J+1}^{(k)}$ from a multivariate normal distribution based on the kth posterior parameter set $\{\boldsymbol{\theta}_{J+1}^{(k)}, \boldsymbol{\Sigma}_{J+1}^{(k)}\}$, with $\xi_{J+1}^{(k)}$ being input using an averaged values estimated for previous quarters up to the Jth quarter.

2. Simulate $N_{i,J+1}^{(k)}$, the number of incidents for the ith combination and $(J+1)$th quarter, from the frequency distribution based the kth set of posterior parameters for frequency model, $\boldsymbol{\beta}_{J+1}^{(k)}$, $\boldsymbol{b}_{J+1}^{(k)}$ and $\xi_{J+1}^{(k)}$.

3. Simulate loss amounts (the number of data breached due to cyber attack incidents) based on number of incidents simulated in Step 2 from the severity distribution for the $(J+1)$th quarter, denoted as $Y_{i,l,J+1}^{(k)}$, $l = 1, 2, \ldots, N_{i,J+1}^{(k)}$ and $i = 1, 2, \ldots, I$, using the estimated parameters $\boldsymbol{\alpha}^T, \mathcal{W}^T, \mathcal{B}^T, \boldsymbol{\gamma}^T$ of severity distribution based on $\mathbf{D}_J$ described in Section 4.3.

4. Calculate $S_{1,J+1}^{(k)}, S_{2,J+1}^{(k)}, \ldots, S_{I,J+1}^{(k)}$, the $(J+1)$th quarter aggregate loss amounts for the respective combination of total I combinations.

In the following we provide some technical/computational notes for the steps stated above for simulating the total (aggregate) loss amounts.

In Step 1, the values of $\boldsymbol{\theta}_{J+1}^{(k)}$ and $\boldsymbol{\Sigma}_{J+1}^{(k)}$ are sampled from their conditional posterior distributions, which are given by Gibbs sampler algorithm; iteration labeling has been altered to fit current MCMC process. One can refer to Section 3.3.1 for complete logistics of the following details:

- Sample $\boldsymbol{\theta}_{J+1}^{(k)}$ from full conditional distribution (3.4)

  (i) compute $\boldsymbol{\mu}_J^{(k-1)}$ and $\Lambda_J^{(k-1)}$ from $\left\{\boldsymbol{\Sigma}_J^{(k-1)}, \boldsymbol{\beta}_1^{(k-1)}, ..., \boldsymbol{\beta}_J^{(k-1)}\right\}$, where

$$\boldsymbol{\mu}_J^{(k-1)} = (\Lambda_0^{-1} + J(\boldsymbol{\Sigma}_J^{(k-1)})^{-1})^{-1}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + J(\boldsymbol{\Sigma}_J^{(k-1)})^{-1}\bar{\boldsymbol{\beta}}_J^{(k-1)}),$$
$$\Lambda_J^{(k-1)} = (\Lambda_0^{-1} + J(\boldsymbol{\Sigma}_J^{(k-1)})^{-1})^{-1};$$

(ii) sample $\boldsymbol{\theta}_{J+1}^{(k)} \sim \mathcal{N}\left(\boldsymbol{\mu}_J^{(k-1)}, \Lambda_J^{(k-1)}\right)$.

- Sample $\boldsymbol{\Sigma}_{J+1}^{(k)}$ from full conditional distribution (3.5)

  (i) compute $\boldsymbol{S}_{\boldsymbol{\theta}}^{(k-1)}$ from $\{\boldsymbol{\theta}_{J+1}^{(k)}, \boldsymbol{\beta}_1^{(k-1)}, ..., \boldsymbol{\beta}_J^{(k-1)}\}$, where

$$\boldsymbol{S}_{\boldsymbol{\theta}}^{(k-1)} = \sum_{j=1}^{J} \left(\boldsymbol{\beta}_j^{(k-1)} - \boldsymbol{\theta}_{J+1}^{(k)})(\boldsymbol{\beta}_j^{(k-1)} - \boldsymbol{\theta}_{J+1}^{(k)}\right)^T;$$

  (ii) sample $\boldsymbol{\Sigma}_{J+1}^{(k)} \sim \mathcal{W}^{-1}\left(\nu_0 + J, \left[\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}}^{(k-1)}\right]^{-1}\right)$.

The value of $\xi_{J+1}^{(k)}$ is input as the average value 2.547 of previous generated 69 ξ_js from their corresponding quarters calculated by (3.6). As discussed in Section 3.2.2, ξ_j is the dispersion parameter for jth quarter and is estimated outside the M-H steps to be entered into GLMM model as a constant. Figure 5.1 plots the 69 quarters' estimated dispersion parameter values, where there is no obvious pattern or trend among those values over time.

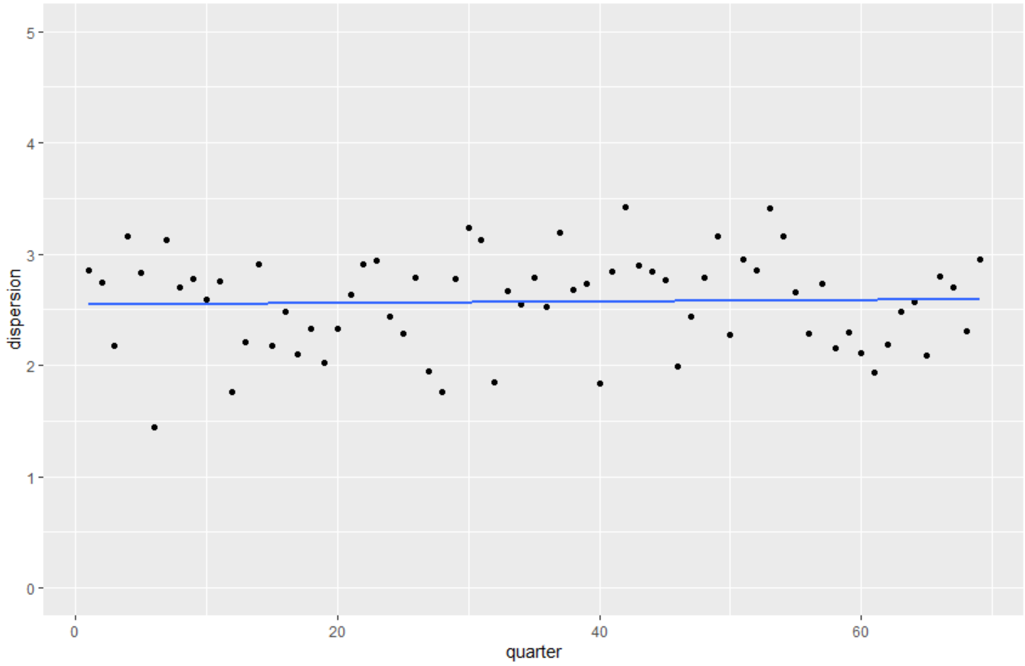


Figure 5.1: Proportion of Zeros

In Step 2, the $\boldsymbol{b}_{J+1}^{(k)}$ are assigned as overall mean value from frequency posterior MCMC analyses in Table 3.2. Per previous GLMM assumption in Section 3.1 that $\boldsymbol{b}$ represents polynomial time effect coefficients that only related to number of quarter labeling, where no combination variance being explained.

Retain last 200 iterated posterior estimations of $\boldsymbol{\theta}_{J+1}^{(k)}$ and $\boldsymbol{\Sigma}_{J+1}^{(k)}$, along with $\xi_{J+1}^{(k)}$, in Step 3, we sample from their corresponding loss frequency posterior distributions given by

$$f(y_{i,J+1}|\mu_{i,J+1},\xi_{J+1}) = \frac{\Gamma(y_{i,J+1}+\xi_{J+1}^{-1})}{\Gamma(\xi_{J+1}^{-1})\Gamma(y_{i,J+1}+1)} \left(\frac{1}{1+\mu_{i,J+1}\xi_{J+1}}\right)^{\xi_{J+1}^{-1}} \left(\frac{\mu_{i,J+1}\xi_{J+1}}{1+\mu_{i,J+1}\xi_{J+1}}\right)^{y_{i,J+1}},$$

where $\mu_{i,J+1}$ is the mean of $Y_{i,J+1}$ such that $\ln(\mu_{i,J+1}) = \eta_{i,J+1} = \boldsymbol{x}_{i,J+1}^T\boldsymbol{\beta}_{J+1} + \boldsymbol{z}_{i,J+1}^T\boldsymbol{b}$, $\boldsymbol{x}_{i,J+1}^T$ and $\boldsymbol{z}_{i,J+1}^T$ are ith covariate combination, and $\boldsymbol{\beta}_{J+1}$ and $\boldsymbol{b}$ are coefficients for fixed effect and random effect.

In Step 4, we sample from the conditional loss severity distribution

$$\begin{aligned}
&f_Y(y;\boldsymbol{\alpha},\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{\gamma},\kappa,\boldsymbol{x})\\
&= p_1(\boldsymbol{\alpha},\boldsymbol{x})\mathbf{1}\{y=0\}\\
&\quad + p_2(\boldsymbol{\alpha},\boldsymbol{x})\frac{f_M(y;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x})}{F_M(c;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x}) - F_M(0^+;\mathcal{W},\mathcal{B},\boldsymbol{\phi},\boldsymbol{x})}\mathbf{1}\{0<y\le c\}\\
&\quad + [1-p_1(\boldsymbol{\alpha},\boldsymbol{x})-p_2(\boldsymbol{\alpha},\boldsymbol{x})]\frac{f_T(y;\boldsymbol{\gamma},\kappa,\boldsymbol{x})}{1-F_T(c;\boldsymbol{\gamma},\kappa,\boldsymbol{x})}\mathbf{1}\{y>c\},
\end{aligned}$$

where $\{p_1,p_2\}\in(0,1)$ are the splicing weights, c is the splicing point which is the threshold separating the moderate and extreme loss values, $\boldsymbol{\alpha}$ is covariate coefficients of zero-inflated weight, $\mathcal{W}$, $\mathcal{B}$ and $\boldsymbol{\phi}$ are parameter vectors of the density of body f_M which is a finite mixture model, and $\boldsymbol{\gamma}$ and κ are parameters of the density of tail f_T. Here we follow the Section 4.3 result and use the Lognormal-Weibull mixture for the body and Pareto for the tail of the distribution given by (4.21).

For the total aggregate loss amount of all types of entities and business within the U.S., repeat above steps by changing coefficients to corresponding covariates related to particular risk combination and obtain 200 ($K = 200$) compound loss series for each of the combinations. When yearly estimates are focused on, repeat iterations for four quarters, $(J+1)$th,..., $(J+4)$th, and obtain the yearly estimates by adding up four consecutive quarters' simulated compound losses. The metrics can then be obtained accordingly.

When simulating loss frequencies and amounts, a common set of covariates is used with each risk combination being assumed to have three normal outcomes: three categorical levels of type of organizations (Medical, Business and Non-business), three categorical levels of type of breaches (External Malicious, Internal Malicious and Internal Negligent) and four geographical regions (Northeast, Midwest, South and West). The detailed classifications of the former two covariates can be found in Table 2.2. These levels can be treated with binary variable in regression modeling with '1' representing that level in the combination and '0' for the rest of other levels.

Using the algorithms described above, we simulate 200 quarterly aggregate loss amounts for each of the 36 combinations. Table 5.2 lists summary statistics of MCMC quarterly level aggregated number of records (total loss) incurred due to breach incidents. For each combination of entity, breach type and region, we list its empirical mean and empirical median. The sample means range from about 400 (B-IN-M) to 24 million (B-EM-S), showing a significant difference in cyber losses with respect to their risk characteristics. The mean excesses median for all the risk combinations, indicating positive skewed loss distributions with a longer or fatter right tail. Furthermore, we name each combination with simplified labels so that remaining tables remain consistency for reference purpose.

Figures 5.2 and 5.3 show box plots of means and medians of simulated aggregate loss distribution in entity, type of breach and location perspective. Based on these plots and results showed in Table 5.2, we have the following observations, which may provide some practical insights for insurance companies developing their cyber insurance products and mitigating cyber insurance risks.

1. **Regional Disparities**: The West region external malicious in business entity demonstrates the highest amount of loss, with median value reaching 9 million, followed by South region external malicious with $835,283$ median loss amount in medical entity, while the Northeast and Midwest regions internal negligent in business entities have the smallest median values of 348 and 440, respectively.

2. **Breach Type Analysis**: External malicious type of breach exhibits the significantly large magnitude in business entity, with median values reaching 9 million in West region and $110,441$ in Northeast region, compared with that of the other breach types. Besides, both external and internal malicious types of breach show the notably large magnitude in medical entity.

3. **Business vs. Non-Business Entities**: Overall, business entities with external malicious breach type experience higher amount of extreme aggregate loss compared to that of non-business entities. This is evidenced by the magnitude of the difference between their means and medians; those of the former are much larger than those of the latter.

4. **Internal Negligent Breaches**: While fewer in amount of loss compared to that of other breach types in general, internal negligent breaches still pose a significant concern, particularly in the South and West regions. Regular training sessions and awareness programs can significantly reduce the likelihood of negligence.

5. **Medical Sector Vulnerability**: The medical sector, particularly in the South and West regions, experiences substantial breaches across all breach types. For instance, the median value for external malicious breaches in the South for medical entities is $835,283$, and for internal malicious in the West region, it is $550,914$.

Entity	Breach	Region	Label	Sample	Sample	Standard
Type	Type			Mean	Median	Deviation
Business	Ext-Malicious	Northeast	B-EM-N	1,100,169	110,441	654,714
		Midwest	B-EM-M	19,448,749	21,353	54,158,543
		South	B-EM-S	23,738,652	22,910	75,234,441
		West	B-EM-W	16,057,435	9,086,214	21,087,536
	Int-Malicious	Northeast	B-IM-N	536,250	319,939	687,189
		Midwest	B-IM-M	8,487	1,570	14,220
		South	B-IM-S	62,776	7,820	162,452
		West	B-IM-W	18,511	450	36,216
	Int-Negligent	Northeast	B-IN-N	119,265	348	278,172
		Midwest	B-IN-M	4,003	440	8,691
		South	B-IN-S	19,793	1,879	37,931
		West	B-IN-W	7,054	2,457	9,520
Non-Bus	Ext-Malicious	Northeast	N-EM-N	177,501	18,480	390,888
		Midwest	N-EM-M	382,931	103,717	542,253
		South	N-EM-S	272,131	63,000	399,549
		West	N-EM-W	375,909	75,638	671,970
	Int-Malicious	Northeast	N-IM-N	29,781	2,064	65,327
		Midwest	N-IM-M	501,588	41,437	1,241,701
		South	N-IM-S	117,679	78,842	124,505
		West	N-IM-W	394,787	104,050	576,395
	Int-Negligent	Northeast	N-IN-N	31,054	2,598	55,452
		Midwest	N-IN-M	13,128	3,789	24,467
		South	N-IN-S	385,358	3,434	1,142,818
		West	N-IN-W	50,099	13,600	104,845
Medical	Ext-Malicious	Northeast	M-EM-N	1,383,445	57,208	3,070,358
		Midwest	M-EM-M	616,941	98,733	1,330,668
		South	M-EM-S	1,841,921	835,283	2,091,217
		West	M-EM-W	2,399,966	425,972	5,372,229
	Int-Malicious	Northeast	M-IM-N	1,019,424	181,570	1,414,608
		Midwest	M-IM-M	241,269	231,804	251,349
		South	M-IM-S	784,253	155,009	1,247,412
		West	M-IM-W	1,269,604	550,914	2,260,585
	Int-Negligent	Northeast	M-IN-N	129,417	65,840	164,314
		Midwest	M-IN-M	162,968	120,613	133,500
		South	M-IN-S	410,179	141,320	611,705
		West	M-IN-W	179,202	87,749	243,482

Table 5.2: Aggregation Statistics

6. **Median vs. Mean Analysis**: Instances where the mean values are significantly higher than the median values, as shown in Figure 5.4, indicate that the loss distribution is right skewed with a heavy tail caused by extreme values. For example, in the South region, the mean value for external malicious breaches in the business category is $23,738,652$, whereas the median value at $22,910$ is nearly 1000 times lower than the mean value, suggesting the presence of extreme loss amounts.

7. **Standard Deviations**: The standard deviation is greater than its corresponding mean for most of the risk combinations. Its significant magnitude relative to the mean suggests a high degree of variability of aggregate losses, likely driven by the presence of extreme loss values, which could have a considerable impact on the overall spread of the losses, such as external malicious activities in business entities in South region having a standard deviation that is 3.17 times of its mean.

8. **Potential Risk Areas**: Regions with consistently high mean and median values across various breach types, such as the West for business entities external malicious, could be identified as potential high-risk areas requiring closer attention and enhanced security measures. Further examination of the loss decomposition may be necessary to determine whether the loss is primarily driven by frequency or severity.

By incorporating numerical results, these insights provide a more detailed understanding of the breach landscape, allowing for more informed decision-making and targeted risk mitigation strategies.

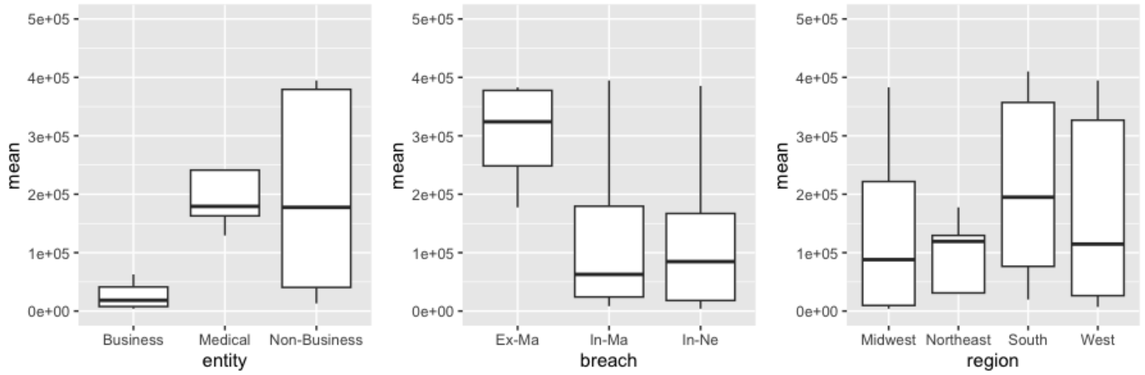


Figure 5.2: Compound Distribution Mean Spread

Besides studying the simulated compound loss distributions, we examine the stationary performance of the MCMC algorithm by looking at overall means range of 100 MCMC simulations. Table 5.3 lays out 95% intervals of 100 means and the corresponding sample mean as shown in Table 5.2) for all the risk combinations. By validating that the interval covers the sample mean effectively, we conclude that the MCMC algorithm used converges satisfactorily.

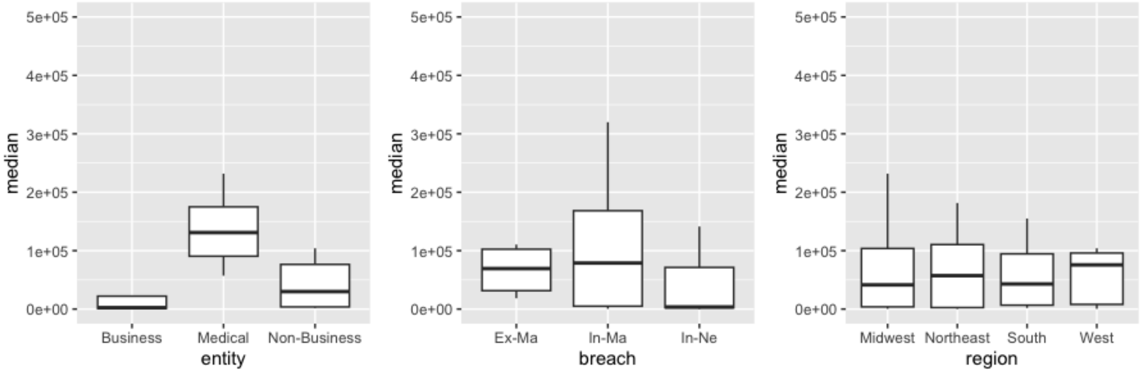


Figure 5.3: Compound Distribution Median Spread

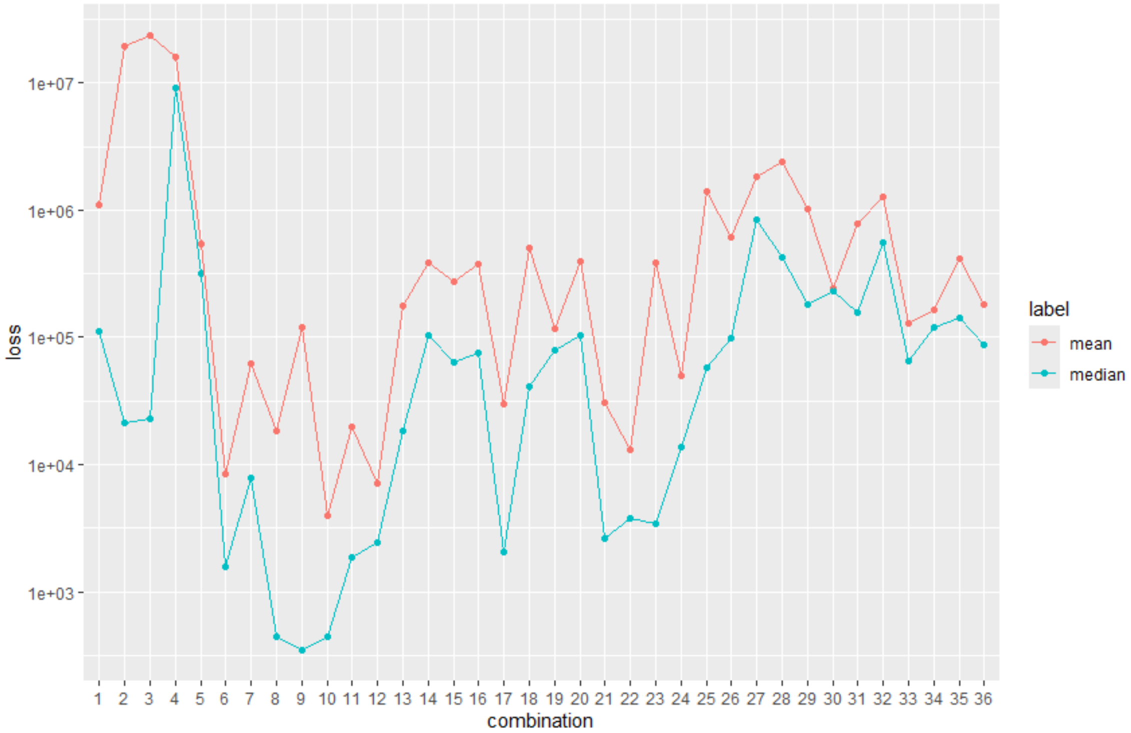


Figure 5.4: Mean Median Comparison

Label	Zeros %	Empirical Mean	95% Interval	VaR$_{95\%}$	CTE$_{95\%}$
B-EM-N	35.7%	1,100,169	(1,089,257, 1,111,081)	3,328,180	6,006,304
B-EM-M	38.4%	19,448,738	(19,279,508, 19,617,997)	76,303,250	150,106,092
B-EM-S	21.4%	23,738,652	(23,396,677, 24,080,621)	92,260,194	250,489,914
B-EM-W	53.9%	16,057,436	(15,881,706, 16,233,165)	36,245,376	53,107,904
B-IM-N	66.8%	536,250	(527,660, 544,834)	974,289	1,505,122
B-IM-M	15.4%	8,486	(8,423, 8,552)	32,506	43,765
B-IM-S	35.3%	62,775	(62,036 63,514)	179,539	546,563
B-IM-W	50.1%	18,512	(18,284, 18,737)	50,347	103,636
B-IN-N	46.2%	119,266	(117,278, 121,252)	351,110	745,718
B-IN-M	36.4%	4,003	(3,941, 4,066)	13,050	23,604
B-IN-S	42.9%	19,793	(19,556, 20,030)	55,886	110,399
B-IN-W	70.1%	7,054	(6,894, 7,212)	11,005	18,163
N-EM-N	16.7%	177,501	(175,546, 179,455)	702,900	1,267,808
N-EM-M	7.7%	382,932	(379,058, 386,804)	1,226,134	1,329,938
N-EM-S	21.4%	272,131	(270,316, 273,947)	1,008,245	1,062,156
N-EM-W	57.1%	375,908	(370,309, 381,509)	852,205	1,712,608
N-IM-N	45.5%	29,781	(29,236, 30,325)	87,091	162,852
N-IM-M	9.3%	501,587	(492,719, 510,457)	2,340,422	3,316,916
N-IM-S	12.7%	117,678	(116,901, 118,457)	296,987	323,067
N-IM-W	27.3%	394,787	(391,18, 398,389)	1,273,052	1,545,840
N-IN-N	53.3%	31,054	(30,657, 31,450)	85,255	147,393
N-IN-M	4.4%	13,128	(12,992, 13,264)	54,572	75,508
N-IN-S	28.6%	385,357	(379,644, 391,073)	1,376,870	3,634,809
N-IN-W	47.2%	50,099	(49,443, 50,753)	118,185	307,356
M-EM-N	26.7%	1,383,445	(1,369,489, 1,397,401)	5,674,011	10,011,968
M-EM-M	30.9%	616,941	(609,548, 624,334)	2,130,852	4,105,648
M-EM-S	23.1%	1,841,920	(1,831,465, 1,852,377)	4,874,641	5,481,983
M-EM-W	27.3%	2,399,966	(2,366,387, 2,433,542)	8,414,694	15,669,185
M-IM-N	30.8%	1,019,425	(1,011,565, 1,027,283)	3,304,454	3,793,125
M-IM-M	14.7%	241,269	(240,222, 242,316)	620,702	858,952
M-IM-S	17.8%	784,253	(780,095, 788,411)	3,038,297	4,347,882
M-IM-W	40.4%	1,269,604	(1,257,046, 1,282,163)	2,921,599	7,229,654
M-IN-N	31.3%	129,417	(128,670, 130,163)	413,678	501,881
M-IN-M	7.1%	162,968	(162,227, 163,709)	340,089	343,315
M-IN-S	14.3%	410,178	(407,630, 412,727)	1,357,308	2,050,971
M-IN-W	42.9%	179,202	(178,188, 180,217)	566,255	781,405

Table 5.3: Risk Measure Statistics

Table 5.3 also shows the proportions of zeros, values at risk at 95% confidence level (VaR$_{95\%}$) and the conditional tail expectations 95% confidence level (CTE$_{95\%}$) for all 36 risk combinations. While VaR$_{95\%}$ represents a quantile for 5% extreme case, the CTE$_{95\%}$ estimates the expected loss if that worst case scenario happens. Upon observing this table's magnitude and risk measures, together with Figures 5.5 and 5.6 which are the dislocations of proportion of zeros and CTE across 36 combinations, we have the following insightful findings.

1. **Proportion of Zeros**

   Across entity types, breach types, and regions, we observe varying proportions of zero incurred loss due to data breaches. Comparing these proportions reveals the combined risk characteristics that have the highest or lowest frequencies of zero-loss occurrences.

   - In the "Business" category, breaches categorized as "Internal Negligent" have the highest proportion of zeros in the Midwest region (70.1%), followed by breaches categorized as "Internal Malicious" that have the second highest proportion in the Northeast region (66.7%).
   - In the "Non-Business" category, breaches categorized as "External Malicious" have the highest proportion of zeros in the West region (57.1%), while breaches categorized as "Internal Negligent" have the lowest proportion in the Midwest region (4.4%).
   - Compared with "Business" and "Non-business" entities, medical entity has a moderate level of zero loss incurred breaches with proportions ranging from 10% to 40%.

2. **Model Performance**

   In terms of predictability level of a model, the width of 95% MCMC simulated interval (fourth column) contains its empirical mean (third column) for all the combinations. It yields a high precise level of and low uncertain estimates.

3. **Value at Risk (VaR)**

   Across different combinations of entity types, breach types, and regions, we observe varying VaR values at 95% confidence level, indicating differences in potential losses at this specified confidence level. Comparing VaR values helps identify which combinations have the highest or lowest potential losses.

   - In the "Business" category, breaches categorized as "External Malicious" have much higher $\text{VaR}_{95\%}$ compared to "Internal Negligent" and "Internal Malicious", suggesting a greater need for cyber security measures targeting external threats in business entities.
   - In the "Business" category, breaches categorized as "External Malicious" in the South region have the highest $\text{VaR}_{95\%}$ (92 million), indicating the greatest extreme loss.
   - The breaches categorized as "Internal Negligent" in the West region within the "Business" category, conversely, have the lowest $\text{VaR}_{95\%}$, indicating that 5% catastrophic loss is no more than 11 thousands.

4. **Conditional Tail Expectation (CTE)**:

   CTE values provide insights into the expected losses beyond the VaR threshold, highlighting the potential scale of extreme breach events.

   - The business external malicious breaches in the South has the highest $\text{CTE}_{95\%}$ of 250 million, indicating the expected losses beyond its $\text{VaR}_{95\%}$ threshold.
   - The $\text{CTE}_{95\%}$ of $3,634,809$ for the non-business internal negligent breaches in the South is significantly higher compared to other regions, suggesting a potentially higher impact of extreme breach events in South region.

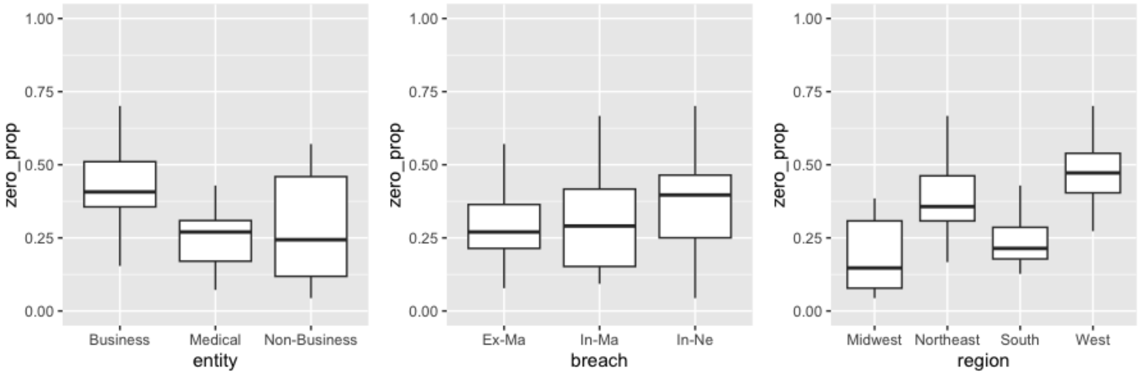


Figure 5.5: Proportion of Zeros

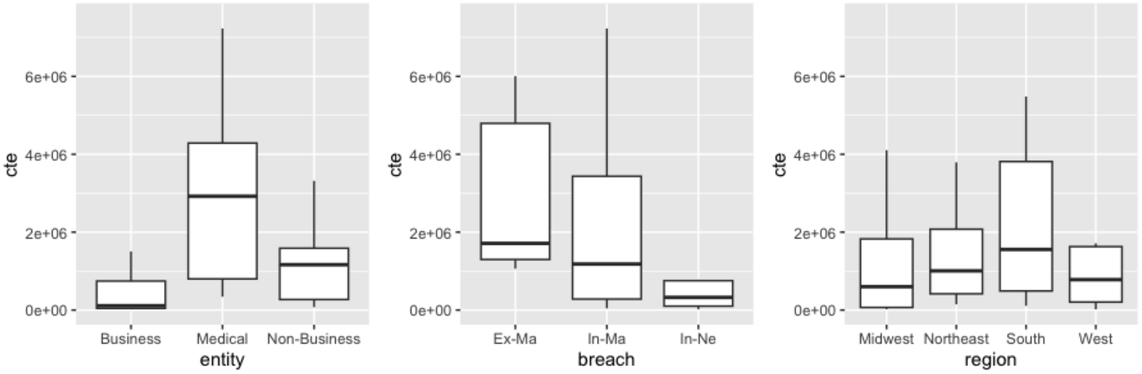


Figure 5.6: Conditional Tail Expectation

5.4 Deductibles, Policy Limits and Reinsurance

This section aims to develop a comprehensive model for estimating insurance losses, providing a range of deductibles and policy limits based on Loss Elimination Ratios (LER). It also seeks to calculate insolvency probabilities using linear loading assumption, compute Increased Limit Factors (ILF), and apply a severity approach to pricing excess-of-loss layers.

We use loss severity distribution selected in Chapter 4 to study the effect of insurance modified coverage of imposing deductibles and policy limits on expected loss cost X for the $(J+1)$th quarter. The LER quantifies the reduction in expected loss for an insurer issuing a policy with a deductible and/or policy limit, compared to the expected loss for an insurer offering full-coverage policies.

Let Y be the loss amount incurred. Deductible of a policy, denoted by d, is the loss retained by the insured, whereas the loss covered by the insurer and paid as claim, denoted by W, can be expressed as

$$W = (Y-d)_+ = \begin{cases} 0, & Y < d, \\ Y-d, & Y \geq d. \end{cases} \tag{5.6}$$

Insurer's LER with deductible d is then defined as

$$\mathrm{LER}_d = \frac{E[Y]-E[W]}{E[Y]} = \frac{E[Y \wedge d]}{E[Y]},$$

where $E[Y \wedge d]$, called the limited expected value, is given by

$$E[Y \wedge d] = \int_0^d y f(y) dy + d \int_d^\infty f(y) dy,$$

and

$$E[Y] = \int_0^\infty y f(y) dy = \int_0^\infty \bar{F}(y) dy,$$

where $\bar{F}(y) = 1 - F(y)$ is called the survival function of cdf F.

Policy limit u is the maximum loss covered by the insurer. The insurer's payment in this case, denoted as L, is given by

$$L = (Y \wedge u) = \begin{cases} Y, & Y < u, \\ u, & Y \geq u. \end{cases}$$

Similar as deductible, insurer's LER with a policy limit l can be written as

$$\mathrm{LER}_u = \frac{E[Y]-E[L]}{E[Y]} = \frac{E[Y]-E[Y \wedge u]}{E[Y]}.$$

To examine the impact of the different values of deductible and the policy limits to the LERs, we consider two sets of deductible values, $d_1 = 300$ and $d_2 = 1,000$, and two policy limits $u_1 = 25,000$ and $u_2 = 75,000$. The corresponding LERs are calculated and displayed in Table 5.4. The amount of deductibles and policy limits are selected for explanation purposes; they could be adjusted upon data structures and specific applications in practice. In Table 5.4, $\Delta\mathrm{LER}_d$ represents the LER difference between d_2 and d_1, whereas

ΔLER_u represents the LER difference between u_1 and u_2. Under this setting, ΔLER_d could represents the premium discount if increasing deductible from d_1 to d_2; similarly, ΔLER_l represents the premium surcharge if increasing limit from u_1 to u_2. By comparing ΔLERs among different combinations and cross comparing ΔLER_d and ΔLER_u, we can draw the following meaningful findings.

- The combinations having large LER_d values mean that they have large scale of low loss amounts, for instance, having $1,000$ deductible eliminates business entities 96% of losses caused by internal malicious in Midwest area (B-IM-M).
- For the combinations with both LER_{d_1} and LER_{d_2} being nearly 0, almost all of their incurred losses are over $1,000$ with no incidents with small loss amounts, so that whether or not applying deductibles would not eliminate any loss from their original loss distribution.
- Combinations with large ΔLER_d values have a significant amount of loss falls between 300 and $1,000$ levels, for example, nearly 50% of loss is eliminated for business internal negligent in Midwest (B-IN-M) after increasing deductible from 300 to $1,000$.
- A combination with a large value of LER_u indicates that it has nearly no extreme losses exceed the policy limit level. For example, business internal malicious in Midwest area (B-IM-M) has both LER_{u_1} and LER_{u_2} being 1, which means that all losses are eliminated after assigning $25,000$ or $75,000$ policy limits.
- Combinations with nearly 0 LER_u values normally have heavy right tail with extreme loss amounts over the policy limit threshold set, such as business entity external malicious tend to incur total losses that are greater than $25,000$ for all the regions.
- A combination with a large value of ΔLER_u has majority of its extreme loss amounts fall between u_1 and u_2 levels. In this case, premium needs to be surcharged in order to adjust increased aggregate loss amount if the policy limit increases from 25,000 to 75,000.

In the situation that the inflation impacts equally across all the risk geographic groups, its effects can be digested by adjusting the base rate. The deductibles and policy limits factor need to be recalculated and the excess loss structure needs to be redesign if there exists geographic variation in inflation because in this situation evenly offsetting base rate would not work. The severity distribution needs to be refit to obtain a new sets of coefficients and LERs, followed with updated deductible and limit factors.

In Section 5.3, we have studied several risk measures for our proposed aggregate loss model; here we make further analysis of deductible and policy limit factors generation, and pure premium and reinsurance cost estimation in dollar amount. Pure premium is calculated based on aggregated loss distribution, where the safety loading factor is set to be 0 for

Label	LER$_{d_1}$	LER$_{d_2}$	ΔLER$_d$	LER$_{u_2}$	LER$_{u_1}$	ΔLER$_u$
B-EM-N	0.002	0.118	11.6%	0.012	0.067	5.5%
B-EM-M	0.000	0.002	0.2%	0.000	0.001	0.1%
B-EM-S	0.000	0.000	0.0%	0.000	0.002	0.2%
B-EM-W	0.000	0.000	0.0%	0.002	0.003	0.1%
B-IM-N	0.000	0.000	0.0%	0.047	0.298	25.1%
B-IM-M	0.044	0.960	91.6%	1.000	1.000	0.0%
B-IM-S	0.020	0.129	10.9%	0.201	1.000	79.9%
B-IM-W	0.158	0.631	47.3 %	1.000	1.000	0.0%
B-IN-N	0.006	0.132	12.6%	0.107	1.000	89.3%
B-IN-M	0.249	0.758	50.9%	1.000	1.000	0.0%
B-IN-S	0.177	0.452	27.5%	1.000	1.000	0.0%
B-IN-W	0.000	0.249	24.9%	1.000	1.000	0.0%
N-EM-N	0.002	0.554	55.2%	0.286	0.292	0.6%
N-EM-M	0.002	0.296	29.4%	0.137	0.142	0.5%
N-EM-S	0.002	0.215	21.3%	0.059	0.318	25.9%
N-EM-W	0.000	0.362	36.2%	0.068	0.241	17.3%
N-IM-N	0.022	0.252	23%	1.000	1.000	0.0%
N-IM-M	0.000	0.043	4.3%	0.055	0.066	1.1%
N-IM-S	0.000	0.064	6.4%	0.657	1.000	34.3%
N-IM-W	0.000	0.106	10.6%	0.067	0.194	12.7%
N-IN-N	0.071	0.523	45.2%	1.000	1.000	0.0%
N-IN-M	0.002	0.165	16.3%	1.000	1.000	0.0%
N-IN-S	0.000	0.094	9.4%	0.057	0.057	0.0%
N-IN-W	0.004	0.171	16.7%	0.233	1.000	76.7%
M-EM-N	0.000	0.009	0.9%	0.011	0.091	8.0%
M-EM-M	0.000	0.000	0.0%	0.114	0.114	0.0%
M-EM-S	0.000	0.000	0.0%	0.011	0.063	5.2%
M-EM-W	0.000	0.000	0.0%	0.019	0.080	6.1%
M-IM-N	0.000	0.000	0.0%	0.111	0.111	0.0%
M-IM-M	0.005	0.176	17.1%	0.118	0.703	58.5%
M-IM-S	0.000	0.000	0.0%	0.035	0.175	14%
M-IM-W	0.000	0.000	0.0%	0.022	0.190	16.8%
M-IN-N	0.000	0.000	0.0%	0.378	1.000	62.2%
M-IN-M	0.000	0.000	0.0%	0.343	1.000	65.7%
M-IN-S	0.000	0.000	0.0%	0.086	0.212	12.6%
M-IN-W	0.002	0.036	3.4%	0.373	0.637	26.4%

Note: $d_1 = 300$, $d_2 = 1000$, $l_1 = 25,000$, $l_2 = 75,000$

Table 5.4: Deductibles, Policy Limits and LERs

the purpose of illustrative analysis. Following the common practice across the property and casualty (P&C) industry for its insurance products rating plan, the incurred loss is capped at 1 million dollar. The reinsurance cost is calculated by borrowing equation (5.6) and setting d to be 1 million, and then computing its expectation. Surplus ratio is the percentage of reinsurance cost out of total loss premium (sum of pure premium and reinsurance cost), which represents the portion of reinsurer covers exceeding the insurer's retained limit or surplus share treaty amount. Table 5.5 displays quarterly estimated dollar amount losses, pure premiums, reinsurance costs and surplus ratios for different risk combinations. According to Table 5.5, we have the following findings on three perspectives that may provide insightful guidance for cyber insurance product developments and premium settings.

Comb.	Label	Estimated Loss	Pure Premium	Reinsurance	Surplus Ratio
1	B-EM-N	39,040	592	1,859	0.76
2	B-EM-M	346,384	312	11,936	0.97
3	B-EM-S	403,039	756	15,071	0.95
4	B-EM-W	299,443	463	6,825	0.94
5	B-IM-N	22,611	853	374	0.30
6	B-IM-M	968	145	64	0.31
7	B-IM-S	4,429	435	304	0.41
8	B-IM-W	1,751	222	100	0.31
9	B-IN-N	7,214	620	244	0.28
10	B-IN-M	547	96	58	0.38
11	B-IN-S	1,842	231	102	0.31
12	B-IN-W	841	131	13	0.09
13	N-EM-N	9,759	410	716	0.64
14	N-EM-M	17,505	879	275	0.24
15	N-EM-S	13,503	808	223	0.22
16	N-EM-W	17,261	588	720	0.55
17	N-IM-N	2,513	289	115	0.28
18	N-IM-M	21,491	302	1,519	0.83
19	N-IM-S	7,141	615	9	0.01
20	N-IM-W	17,916	569	675	0.54
21	N-IN-N	2,594	296	88	0.23
22	N-IN-M	1,348	184	91	0.33
23	N-IN-S	17,590	258	1,648	0.86
24	N-IN-W	3,731	385	170	0.31
25	M-EM-N	46,466	715	2,450	0.77
26	M-EM-M	25,153	779	1,247	0.62
27	M-EM-S	57,757	896	1,445	0.62
28	M-EM-W	70,625	1,197	2,761	0.70
29	M-IM-N	36,843	752	1,195	0.61
30	M-IM-M	12,323	913	24	0.03
31	M-IM-S	30,185	934	1,151	0.55
32	M-IM-W	43,530	1,280	1,409	0.52
33	M-IN-N	7,676	648	60	0.08
34	M-IN-M	9,145	736	4	0.01
35	M-IN-S	18,444	953	479	0.33
36	M-IN-W	9,830	775	109	0.12

Table 5.5: Pure Premium Analysis

Pure Premium

- The largest pure premiums are often associated with medical entity malicious types of breaches in West region, where they can reach up to $1,000 per quarter.
- Following closely are medical entities in South region, with pure premiums ranging from $953 for internal negligent breach type and $934 for internal malicious breach type.
- Internal negligence in business entities in Midwest and West regions has overall low pure premium level, with the lowest pure premium $96 in Midwest region, followed by $131 pure premium per quarter in West region.

Reinsurance Cost

- The largest reinsurance costs are observed for external malicious breaches in business entities, particularly in South and Midwest regions which are nearly $12,000 and $15,000 per quarter, respectively.
- Malicious breaches in medical entities also attract thousands of high reinsurance premiums, especially in regions with high healthcare activities such as Northeast and West.
- Reinsurance premiums for internal negligent breaches in medical entities in Midwest tend to be the smallest having only $4 per quarter. Besides, the reinsurance cost for internal malicious breaches in non-business entities in South region is estimated to be $9 per quarter.

Relativity Comparison

- The reinsurance cost is tremendously higher than the pure premium for business entities with external malicious breach activities, with over 70% surplus ratio for all the regions.
- The scale of the reinsurance cost compared to the pure premium for medical entity external malicious activities is also considerably large with over 60% surplus ratios for all the regions.
- Medical entity internal breach activities normally have a low scale of reinsurance cost compared to the pure premium level, especially in Midwest region where the reinsurance cost portion is less than 5%.

Figure 5.6 further shows the magnitude between pure premium and reinsurance cost, where we assign combination number from 1 to 36 to represent 36 combinations of entity, breach type and regions; the corresponding relationship can be referred to Table 5.5 first and second columns.

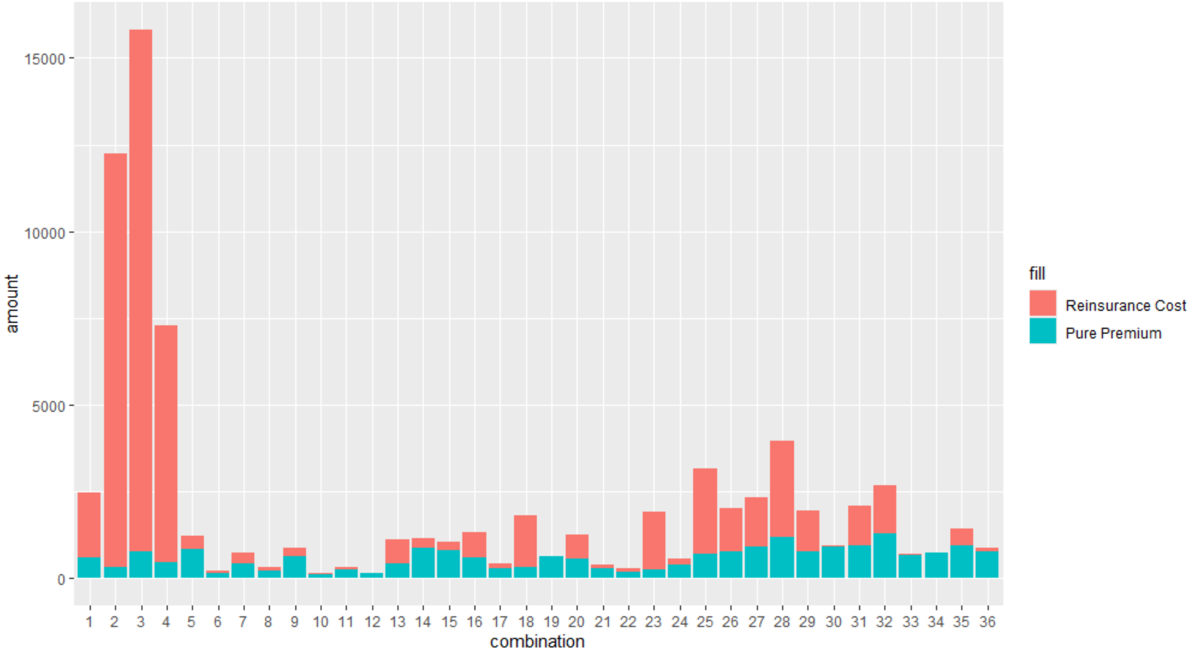


Figure 5.7: Premium Decomposition

Reinsurance arrangements offer primary insurers several key advantages. They enable insurers to maintain prudent risk levels by transferring significant exposures to another company. Additionally, they empower insurers to meet client demands for broader coverage by leveraging the financial resources of a reinsurer. Moreover, primary insurers gain access to the valuable underwriting expertise, experience, and claims handling capabilities of the reinsurer. These factors are critical for insurers seeking to expand their operations and mitigate loss exposure, particularly in countries with a large number of insurers and limited resources. Consider XYZ Insurance, which provides cyber insurance to businesses. By partnering with ABC Reinsurance, XYZ can transfer a portion of its risk, ensuring financial stability during major cyber incidents. This allows XYZ to offer broader coverage to clients while benefiting from ABC's underwriting and claims management expertise, supporting growth and risk mitigation.

5.5 Application to Current U.S. Cyber Insurance Development

In this section, we discuss the cyber insurance premium development from the model structure perspective and the product design perspective. Treatments in pricing cyber insurance products are similar to that for property insurance, which first estimates compound loss distribution and coefficient factors sequentially based on the relevant data. However, unlike other property insurance products that have large enough book of policies and records accumulated over time, digital property remains a very thin level of repository. Since the Internet development history is relatively short, histories of data breach incidents and related losses are not extensive or recorded with low documentation quality. The repositories of information security breaches do not sufficient enough due to limited years of accumulation and firms' unwillingness to reveal incurred breach details. The insurance companies have no standard scoring or actuarial tables to make pricing determinations for cyber risk policies, which requires necessary methodological innovations and updates. Given this fact, some of the risk groups have been disqualified under current rating system, our proposed macro-level modeling approach can solve this problem by enabling modeler to find information leakage and improve prediction power. Once advanced risk segmentation models are developed, information leakage and profit margin can be updated and perfected accordingly.

Utilizing risk characteristics, our model divides homogeneous risks into segments. Then, product designers can decide whether to implement policy limits or seek reinsurance. It can be broken down into four steps: establishing the base rate, multiplying risk factors, applying discount of deductible or surcharge of policy limit modifications, and factoring in expense retention. The first and second steps lay the groundwork for the entire pricing process. They are frequently carried out using an experience-based pricing techniques with preliminary data for analysis. The four steps can be described below.

- **Establishing the Base Rate**
  It calculates a base premium rate for cyber insurance coverage, for example, $1,000 quarterly, based on industry averages and historical data for small to medium-sized enterprises (SMEs).

- **Multiplying Risk Factors**
  It identifies risk factors such as the type of industry, covered breach types, and geographical location. For instance, a company in the business sector with some level of cyber protections in place might receive a risk factor multiplier of 1.2[1], while a company in the non-business sector with minimal security might have a multiplier of 1.0 (base level factor). If an SME in retail has a base rate of $1,000, the adjusted premium for that business company would be $1,000 × 1.2 = $1,200.

- **Applying Discounts or Surcharges**
  It applies adjustments based on policy features. If the business opts for a higher deductible, such as $1,000, it might receive a discount of 5%, reducing the premium to $1,200 × 0.95 = $1,140. Conversely, if the business requests higher policy limits, it might incur a surcharge of 10%, increasing the premium to $1,200 × 1.1 = $1,320.

[1] A relativity factor for business entities taking non-business entities as base level.

- **Factoring in Expense Retention**
  It includes an expense retention factor to cover administrative and operational costs. If the expense retention factor is 5%, it adds \$1,200 × 0.05 = \$60 to the final premium, resulting in a total premium of \$1,260 for the insured SME.

Our cyber risk loss aggregation results are carried out within a Bayesian framework, which proves to be a useful and effective prediction tool for estimating future loss among segmentation with confidence. The calculations for those risk measures in Table 5.3 can be used for both internal risk management purpose and for regulatory capital, the capital requirements set by the insurance supervisors, purpose. The quantitative insights of deductibles, limits and reinsurance in Section 5.4 provide relative flexible rate adjustments information when setting manual rates in premium pricing. Insurance companies are able to maintain high solvency in the differentiated pricing case compared to the case of non-differentiated pricing (Pal et al., 2017). Nevertheless, given the high uncertainty of cyber risk quantification, frequently monitoring external force is necessary.

Besides modeling structure limitation, the amount of cyber insurance product exposures is accumulating year by year due to the nature that digital assets is a key component of business operations. In contrast to homeowner insurance that the replacement cost diminishes over time due to depreciation, the cyber insurance policy determines value by taking the cost to replace/restore digital assets and the cost increases in time due to inflation of the value in digital asset ecosystem. Understanding this mosaic is essential, as is the differentiation between cyber risks and general property risks, including their corresponding attack exposures. It leads to a paradigm shift in the insurance industry, where traditional models no longer suffice in adequately protecting businesses from the complexities of cyber threats. The rise in cyber insurance product exposures underscores the growing recognition among businesses of the need for specialized coverage tailored to the unique challenges posed by digital assets.

The problem of adverse selections is another factor that needs to be considered when designing cyber risk insurance policies. In the absence of perfect information, the competitive outcome in markets for insurance may be non-optimal not only compared to the infeasible optimum that would have occurred if information were perfect but also compared to optima that are feasible (Pauly, 1978). It happens when the insurance purchaser has control over actions in the present that affect the future state of nature but in which the insurer cannot directly observe the insured's actions. For cyber risk insurance, the adverse selection is mainly about the likelihood of a security breach and the transparency of the amount of digital assets. For example, an organization that has more hidden exposures of cyber activities would be more prone to purchase cyber insurance than an organization with average exposed to cyber risks. In order to tackle this issue, insurers could require an information security audit before issuing a policy. Another meaningful response to the adverse selection problem is to utilize our raised methodologies in previous chapters by segmenting high-risk users into different risk groups through examining their losses exhibition especially for tail behaviour, and accurately assigning the premium for those groups. Additional solutions can be that underwriters require full exposure information of digital assets and require a waiting period to pass before the policy is effective. Moreover, as the digital landscape continues to evolve, so do the tactics employed by cybercriminals. This necessitates a proactive approach to risk management, wherein businesses not only invest in robust cybersecurity measures but also leverage cyber insurance as a vital component of their risk mitigation strategy. A cyber insurance policy can provide financial protection against the potential costs associated with cyber incidents, and it can also facilitate access to resources for incident response, recovery, and post-incident support. Furthermore, the dynamic nature of cyber risks requires constant reassessment and adaptation of insurance practices. Insurers must stay abreast of emerging threats and evolving regulatory landscapes to provide comprehensive coverage that addresses the evolving needs of businesses. This includes offering innovative solutions such as cyber risk assessments, threat intelligence, and cybersecurity training to help businesses enhance their cyber resilience.

In essence, the increasing prominence of cyber insurance signifies a fundamental shift in how businesses perceive and manage cyber risks. By embracing specialized insurance solutions and adopting a proactive approach to cybersecurity, businesses can navigate the digital landscape with confidence, knowing they have the necessary safeguards in place to protect their most valuable assets. Our focus is on the cyber risks quantification and mitigation that modelers working with digital asset-related insurance products would typically prioritize for monitoring and analysis. By developing above robust investigations of where the potential updates and improvements reside, insurers can methodically discern their frameworks best tailored to address them. Furthermore, this knowledge becomes instrumental in determining the most fitting risk transfer mechanisms - cyber insurance, paving the way for a future where the cyber community is both innovative and secure for the everyday user.

Chapter 6

Conclusion

The thesis first describes the PRC chronology data, including preliminary analysis with descriptive statistics, exploratory analysis of utilized features, and cluster analysis of geographical information. It then proposes a Bayesian negative binomial GLMM for quarterly cyber incidents recorded by the PRC dataset, capturing within-quarter heterogeneity effects and allowing subject-specific predictions. Following this, a zero-inflated mixture and composite regression model for cyber loss amounts (the number of data breached) is presented, detailing model fitting, selection, and applications from both insurers' and insureds' perspectives. The thesis concludes by proposing approaches for generating aggregate losses and implementing strategies for the insurance industry, and discussing the impact of different deductibles, limits, and reinsurance practices, with applications to the U.S. cyber insurance market. In this chapter, we conclude this thesis by stating the contributions of our research, addressing the limitations encountered, and suggesting avenues for future research to further enhance the field of cyber risk modeling and mitigation.

The contributions of this thesis to the related research areas can be described as follows. In modeling the loss frequency, we investigate the use of average severity as one of the subject-specific covariates via the regression within the generalized linear mixed model (GLMM); thus, the dependence between the frequency and the severity of cyber risks is considered. Meanwhile, we model the time trend effects as a group-specific factor in order to explain the change in data breach incidents over time. Besides examining fixed effects, we adopt the MCMC method to extract random effects on several different explanatory variables. We estimate parameters of GLMM under the NB distribution with a non-constant scale parameter by combining the maximum likelihood estimation with the MCMC method. We add to the existing literature the implementation of our proposed estimation procedure in the actuarial context, which may be of interest to other researchers and practitioners in the related fields. In modeling the loss severity pattern, we propose a zero-inflated mixture composite regression (Zi-MCR) model (3-components spliced distribution). It features a flexible finite mixture model (FMM) with different types of distributions modeling the non-zero body component and an extreme distribution modeling the tail component, and incorporates the rate of point mass at zero, the FMM and the extreme distribution into a GLM structure to fully utilize the risk characteristics by treating them as covariates within the regression framework. Hence, our work enables cyber risks to be completely quantified under one distribution taking into consideration the zero loss component, and positive loss amounts with the heavy-tailed nature. Furthermore, our methodologies provide a meaningful and innovative approach for evaluating aggregate cyber losses, which sets the ground for estimating feature coefficients and generating premium factors.

Cyber risk loss exposures permeate every facet of an organization's operations, making the consequences of a data breach potentially catastrophic. Unlike other kinds of property and casualty insurance risk that capping incurred losses at a 95% level could effectively rule out extreme values, the cyber risk has the nature that, even upon the logarithm, the loss distribution is very heavily skewed to be capped at a bell shaped distribution. The traditional insurance pricing sets up a policy limit and does not consider extreme losses when training the model. However, this technique can not be applied to analyze cyber risks as it is difficult to set such a limit so that cyber losses could be modeled via one single distribution. We bring up a more statistically rigorous attempt to incorporate excess zeros, mixture components and heavy tail of cyber losses in a single and statistically consistent step where other estimation processes, such as covariates dependence, can also be carried on.

An important aspect of this thesis is the use of the publicly available Privacy Rights Clearinghouse (PRC) Data Breach Chronology dataset on developing actuarial approaches to quantify cyber loss frequencies and severities. However, the quality of available data and whether the data represents well cyber risks in general lead to a limitation of this study. The fact that firms do not reveal details concerning security breaches reduces data accuracy, and not voluntarily reporting cyber breaches leads to data inadequacy. Moreover, PRC has stopped updating latest breach incidents since 2019, which causes data inconsistency in a time trend manner. The availability of high-quality data such as policy or claim database in the future would open up new research opportunities. Our model is subjective and can be modified to accommodate the features of new dataset and the purpose of prediction.

Despite the limitations, our study of cyber risks based on the frequency-severity approach is important for insurance companies in mitigating and managing their risks given that the functioning of the insurance business is a complex process. Enterprises need to take several measures in dealing with cyber risks: operations based on statistical modeling in actuarial analysis process, ensuring the balance and adequacy of tariffs in pricing process and adjusting premium rates in insurance marketing. Our research results can be used as a differential indicator on different organization types and geographical locations for developing cyber insurance products. In addition, our study can also be useful for data security officers and scientists, and other potential corporate stakeholders for them to better understand the impact of the cyber risks for business operations.

As previously stated, all of our results are based on the assumption of equal exposure, whereas exposure is the most crucial factor in determining the pure premium. Cyber risk loss exposures are any conditions that present the possibility of financial loss to an organization from property, net income, and liability as a consequence of advanced technology transmissions, operations, maintenance, development, and support. Training the predictive model under the assumption of non-level exposure in a defined time period would be an important direction for future research once prior experience data with exposure information is obtained. This approach will allow for more accurate and reliable premium calculations by accounting for variations in exposure levels. Additionally, incorporating dynamic exposure metrics into the predictive models can further enhance their robustness and applicability to real-world scenarios. Another promising research direction involves developing a precise formula to convert value of digital units into dollar amounts, providing an accurate financial estimate of cyber breach losses. This method could consider factors such as inflation, the organizational structure and size of different companies, and other economic indicators to ensure that the financial impacts are appropriately quantified. By addressing these elements, researchers can create more comprehensive models that reflect the true financial risks associated with cyber breaches.

Bibliography

Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. In *International conference on foundations of data organization and algorithms*, pages 69–84. Springer.

Ahn, S., Kim, J. H., and Ramaswami, V. (2012). A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics*, 51(1):43–52.

Allen, D. E., Singh, A. K., and Powell, R. J. (2013). Evt and tail-risk modelling: Evidence from market indices and volatility series. *The North American Journal of Economics and Finance*, 26:355–369.

Antonio, K. and Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58–76.

Araiza Iturria, C. A., Godin, F., and Mailhot, M. (2021). Tweedie double glm loss triangles with dependence within and across business lines. *European Actuarial Journal*, 11(2):619–653.

Archer, N., Sproule, S., Yuan, Y., Guo, K., and Xiang, J. (2012). *Identity Theft and Fraud: Evaluating and managing risk*. University of Ottawa Press.

Arcidiacono, P. and Jones, J. B. (2003). Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica*, 71(3):933–946.

AXA, E. G. . (2019). Future risks report.

Baumgartner, F. R. and Jones, B. D. (2015). *The politics of information: Problem definition and the course of public policy in America*. University of Chicago Press.

BBC News (2021). US companies hit by 'colossal' cyber-attack, https://www.bbc.com/news/world-us-canada-57703836.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.

Bell, G. and Ebert, M. (2015). Health care and cyber security: Increasing threats require increased capabilities.

Bermúdez, L. and Karlis, D. (2012). A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis*, 56(12):3988–3999.

Bernardi, M., Maruotti, A., and Petrella, L. (2012). Skew mixture models for loss distributions: a bayesian approach. *Insurance: Mathematics and Economics*, 51(3):617–623.

Bessy-Roland, Y., Boumezoued, A., and Hillairet, C. (2021). Multivariate hawkes process for cyber insurance. *Annals of Actuarial Science*, 15(1):14–39.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Blostein, M. and Miljkovic, T. (2019). On modeling left-truncated loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, 85:35–46.

Blough, D. K., Madden, C. W., and Hornbrook, M. C. (1999). Modeling risk using generalized linear models. *Journal of health economics*, 18(2):153–171.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Buch-Larsen, T., Nielsen, J. P., Guillén, M., and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–516.

Bühlmann, H. (1980). An economic premium principle. *ASTIN Bulletin: The Journal of the IAA*, 11(1):52–60.

Bühlmann, H. (2007). *Mathematical methods in risk theory*, volume 172. Springer Science & Business Media.

Carfora, M. F. and Orlando, A. (2019). Quantile based risk measures in cyber security. In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–4. IEEE.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Charpentier, A. and Oulidi, A. (2010). Beta kernel quantile estimators of heavy-tailed loss distributions. *Statistics and computing*, 20(1):35–55.

Chatfield, C. and Collins, A. J. (2018). *Introduction to multivariate analysis*. Routledge.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.

Chiou, Y.-C. and Fu, C. (2015). Modeling crash frequency and severity with spatiotemporal dependence. *Analytic Methods in Accident Research*, 5:43–58.

CIAB (2021). Commercial property/casualty market survey. *The Council of Insurance Agents Brokers.*

CISA (2022). Russia cyber threat overview and advisories.

Clark, D. R. (2022). Alternative to tweedie in pure premium glm. In *CAS E-Forum.*

Clayton, D. G. (1996). Generalized linear mixed models. *Markov chain Monte Carlo in practice*, 1:275–302.

ColonyWest (2023). A history of cyber liability insurance.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Cruz, M. G., Peters, G. W., and Shevchenko, P. V. (2015). *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk.* John Wiley & Sons.

Data Accountability and Trust Act (2019). https://www.congress.gov/bill/116th-congress/house-bill/1282.

Data Security and Breach Notification Act (2015). https://www.congress.gov/bill/114th-congress/house-bill/1770.

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101:485–497.

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

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375.

Edwards, B., Hofmeyr, S., and Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2(1):3–14.

Eling, M. (2020). Cyber risk research in business and actuarial science. *European Actuarial Journal*, pages 1–31.

Eling, M. and Jung, K. (2018). Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance: Mathematics and Economics*, 82:167–180.

Eling, M. and Jung, K. (2022). Heterogeneity in cyber loss severity and its impact on cyber risk measurement. *Risk Management*, 24(4):273–297.

Eling, M. and Loperfido, N. (2017). Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: mathematics and economics*, 75:126–136.

Equifax (2017). Equifax acquires data-crédito.

Everitt, B. (2013). *Finite mixture distributions.* Springer Science & Business Media.

Ewold, F. (1991). Insurance and risk. *The Foucault effect: Studies in governmentality*, 197210:201–202.

Fahrenwaldt, M. A., Weber, S., and Weske, K. (2018). Pricing of cyber insurance contracts in a network model. *ASTIN Bulletin: The Journal of the IAA*, 48(3):1175–1218.

Farkas, S., Lopez, O., and Thomas, M. (2021). Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105.

Farley, J. (2022). The cyber insurance market struggles with continued hardening market conditions.

FBI (2000). Internet Crime Complaint Center (IC3), https://www.fbi.gov/investigate/cyber.

Fitch (2023). Us cyber insurers see favorable premium growth, results in 2023. *FITCH WIRE.*

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Frees, E. W., Lee, G., and Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1):4.

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

Fung, T. C., Badescu, A. L., and Lin, X. S. (2019a). A class of mixture of experts models for general insurance: Application to correlated claim frequencies. *ASTIN Bulletin: The Journal of the IAA*, 49(3):647–688.

Fung, T. C., Badescu, A. L., and Lin, X. S. (2019b). A class of mixture of experts models for general insurance: Theoretical developments. *Insurance: Mathematics and Economics*, 89:111–127.

Fung, T. C., Jeong, H., and Tzougas, G. (2024). Soft splicing model: bridging the gap between composite model and finite mixture model. *Scandinavian Actuarial Journal*, 2024(2):168–197.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* CRC Press.

Gan, G. and Valdez, E. A. (2018). Fat-tailed regression modeling with spliced distributions. *North American Actuarial Journal*, 22(4):554–573.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205–215.

Garrido, J. and Zhou, J. (2009). Full credibility with generalized linear and mixed models. *ASTIN Bulletin: The Journal of the IAA*, 39(1):61–80.

Gartner (2022). Gartner identifies three factors influencing growth in security spending.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Gilks, W. R. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*.

Giudici, P. and Raffinetti, E. (2021). Cyber risk ordering with rank-based statistical models. *AStA Advances in Statistical Analysis*, 105:469–484.

Gordon, L. A. and Loeb, M. P. (2002). The economics of information security investment. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):438–457.

Gui, W., Huang, R., and Lin, X. S. (2018). Fitting the erlang mixture model to data via a gem-cmm algorithm. *Journal of Computational and Applied Mathematics*, 343:189–205.

Hardy, M. R. (2006). An introduction to risk measures for actuarial applications. *SOA Syllabus Study Note*, 19.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64(328):1459–1471.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hathaway, R. J. (1986). Another interpretation of the em algorithm for mixture distributions. *Statistics & probability letters*, 4(2):53–56.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

Hofstetter, H., Dusseldorp, E., Van Empelen, P., and Paulussen, T. W. (2014). A primer on the use of cluster analysis or factor analysis to assess co-occurrence of risk behaviors. *Preventive medicine*, 67:141–146.

Hua, L. (2015). Tail negative dependence and its applications for aggregate loss modeling. *Insurance: Mathematics and Economics*, 61:135–145.

IBM (2020). Cost of a data breach report 2021.

Ispirian, K., Margarian, A., and Zverev, A. (1974). A monte-carlo method for calculation of the distribution of ionization losses. *Nuclear Instruments and Methods*, 117(1):125–129.

Jacobs, J. (2014). Analyzing ponemon cost of data breach. http://datadrivensecurity.info/blog/posts/2014/dec/ponemon/, (accessed 22.09.28).

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jeong, H., Valdez, E. A., Ahn, J. Y., and Park, S. (2021). Generalized linear mixed models for dependent compound risk models. *Variance*, 14(1).

Jevtić, P. and Lanchier, N. (2020). Dynamic structural percolation model of loss distribution for cyber risk of small and medium-sized enterprises for tree-based lan topology. *Insurance: Mathematics and Economics*, 91:209–223.

Joe, H. and Zhu, R. (2005). Generalized poisson distribution: the property of mixture of poisson and comparison with negative binomial distribution. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(2):219–229.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Katz, R. W. (1977). Precipitation as a chain-dependent process. *Journal of Applied Meteorology (1962-1982)*, pages 671–676.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. John Wiley & Sons.

Kshetri, N. (2020). The evolution of cyber-insurance industry and market: An institutional analysis. *Telecommunications Policy*, 44(8):102007.

Lee, G. Y. and Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics*, 87:115–129.

Leswing, K. (2020). Twitter hackers who targeted elon musk and others received $121,000 in bitcoin, analysis shows. *CNBC TECH*.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., and Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mrna processing defects in human genes. *Proceedings of the National Academy of Sciences*, 108(27):11093–11098.

Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., and Liu, M. (2015). Cloudy with a chance of breach: Forecasting cyber security incidents. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 1009–1024.

Maillart, T. and Sornette, D. (2010). Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, 75(3):357–364.

Malavasi, M., Peters, G. W., Shevchenko, P. V., Trück, S., Jang, J., and Sofronov, G. (2021). Cyber risk frequency, severity and insurance viability. *arXiv preprint arXiv:2111.03366.*

Malavasi, M., Peters, G. W., Shevchenko, P. V., Trück, S., Jang, J., and Sofronov, G. (2022). Cyber risk frequency, severity and insurance viability. *Insurance: Mathematics and Economics*, 106:90–114.

Markov, A. A. (1906). Extension of the law of large numbers to dependent quantities. *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)*, 15(1):135–156.

Marriott (2020). Marriott international notifies guests of property system incident. *Marriott International News Center.*

Mazzoccoli, A. and Naldi, M. (2020). Robustness of optimal investment decisions in mixed insurance/investment cyber risk management. *Risk Analysis*, 40(3):550–564.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models.* Routledge.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170.

McCulloch, C. E. (2006). Generalized linear mixed models. *Encyclopedia of Environmetrics*, 2.

McCulloch, C. E. and Searle, S. R. (2004). *Generalized, linear, and mixed models.* John Wiley & Sons.

McLennan, M. (2021). Cyber insurance market overview: Fourth quarter 2021. *Marsh Cyber Risk Report.*

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA*, 27(1):117–137.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341.

Michael A. Bean, FCAS, C. F. F. P. (2020). Exposure measures for pricing and analyzing the risks in cyber insurance.

Miljkovic, T. and Grün, B. (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, 70:387–396.

Mosier, C. I. (1951). I. problems and designs of cross-validation 1. *Educational and Psychological Measurement*, 11(1):5–11.

NAIC (2020). National Association of Insurance Commissioners Report on the Cybersecurity Insurance Market, https://www.insurancejournal.com/app/uploads/2021/11/naic-cyber_insurance-report-2020.pdf.

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

O'Leary, D. P. (1990). Robust regression computation using iteratively reweighted least squares. *SIAM Journal on Matrix Analysis and Applications*, 11(3):466–480.

Pal, R., Golubchik, L., Psounis, K., and Hui, P. (2017). Security pricing as enabler of cyber-insurance a first look at differentiated pricing markets. *IEEE Transactions on Dependable and Secure Computing*, 16(2):358–372.

Park, M. H. and Kim, J. H. (2016). Estimating extreme tail risk measures with generalized pareto distribution. *Computational Statistics & Data Analysis*, 98:91–104.

Pauly, M. V. (1978). Overinsurance and public provision of insurance: The roles of moral hazard and adverse selection. In *Uncertainty in economics*, pages 307–331. Elsevier.

Peel, D. and MacLahlan, G. (2000). Finite mixture models. *John & Sons*.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, pages 863–867.

Pinheiro, J. C. and Chao, E. C. (2006). Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):58–81.

Poudyal, C. (2021). Truncated, censored, and actuarial payment-type moments for robust fitting of a single-parameter pareto distribution. *Journal of Computational and Applied Mathematics*, 388:113310.

PRC (2019). Privacy Rights Clearinghouse Chronology of Data Breaches, https://privacyrights.org/data-breaches.

Quijano Xacur, O. A. et al. (2011). *Property and casualty premiums based on tweedie families of generalized linear models*. PhD thesis, Concordia University.

Quijano Xacur, O. A. and Garrido, J. (2015). Generalised linear models for aggregate claims: to tweedie or not? *European Actuarial Journal*, 5(1):181–202.

Rathee, A. (2020). Data breaches in healthcare: A case study. *CYBERNOMICS*, 2(2):25–29.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics*, 9(1):141–157.

Reynkens, T., Verbelen, R., Beirlant, J., and Antonio, K. (2017). Modelling censored losses using splicing: A global fit strategy with mixed erlang and extreme value distributions. *Insurance: Mathematics and Economics*, 77:65–77.

Roberts, S. J. (1997). Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2):261–272.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471.

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, 2(2):121–135.

Rudolph, M. J. (2022). 15th Annual Survey of Emerging Risks, https://www.casact.org/sites/default/files/2022-08/15th-survey-emerging-risks.pdf.

Sattayatham, P. and Talangtam, T. (2012). Fitting of finite mixture distributions to motor insurance claims. *Journal of Mathematics and Statistics*, 8(1):49–56.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Schnell, W. (2020). Does cyber risk pose a systemic threat to the insurance industry? working paper. https://www.alexandria.unisg.ch/260003/.

Scurrah, K. J., Palmer, L. J., and Burton, P. R. (2000). Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (glmms) and gibbs sampling in bugs. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 19(2):127–148.

Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R., and Ahmad Khan, R. (2020). Healthcare data breaches: insights and implications. In *Healthcare*, volume 8, page 133. MDPI.

Septiany, R., Setiawaty, B., and Purnaba, I. G. P. (2020). The use of monte carlo method to model the aggregate loss distribution. *Al-Jabar: Jurnal Pendidikan Matematika*, 11(1):179–190.

Sheehan, B., Murphy, F., Kia, A. N., and Kiely, R. (2021). A quantitative bow-tie cyber risk classification and assessment framework. *Journal of Risk Research*, 24(12):1619–1638.

Shi, P. (2016). Insurance ratemaking using a copula-based multivariate tweedie model. *Scandinavian Actuarial Journal*, 2016(3):198–215.

Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417–428.

Smith, A. F. and Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Sun, H., Xu, M., and Zhao, P. (2021). Modeling malicious hacking data breach risks. *North American Actuarial Journal*, 25(4):484–502.

Sun, M. and Lu, Y. (2022). A generalized linear mixed model for data breaches and its application in cyber insurance. *Risks*, 10(12):224.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Tse, Y.-K. (2009). *Nonlife actuarial models: theory, methods and evaluation.* Cambridge University Press.

Tzougas, G., Vrontos, S., and Frangos, N. (2014). Optimal bonus-malus systems using finite mixture models. *ASTIN Bulletin: The Journal of the IAA*, 44(2):417–444.

Varonis, F. S. (2021). 2021 financial services data risk report.

Werner, G. and Modlin, C. (2010). Basic ratemaking. In *Casualty Actuarial Society*, volume 4, pages 1–320.

Wheatley, S., Maillart, T., and Sornette, D. (2016). The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, 89(1):1–12.

Williams, P. A. and Woodward, A. J. (2015). Cybersecurity vulnerabilities in medical devices: a complex environment and multifaceted problem. *Medical Devices: Evidence and Research*, pages 305–316.

Wolfinger, R. and O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3-4):233–243.

Xie, X., Lee, C., and Eling, M. (2020). Cyber insurance offering and performance: an analysis of the us cyber insurance market. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 45(4):690–736.

Xu, M., Schweitzer, K. M., Bateman, R. M., and Xu, S. (2018). Modeling and predicting cyber hacking breaches. *IEEE Transactions on Information Forensics and Security*, 13(11):2856–2871.

Yau, K., Yip, K., and Yuen, H. (2003). Modelling repeated insurance claim frequency data using the generalized linear mixed model. *Journal of Applied Statistics*, 30(8):857–865.

Yee, T. W. (2015). *Vector generalized linear and additive models: with an implementation in R*, volume 10. Springer.

Young, D., Lopez Jr, J., Rice, M., Ramsey, B., and McTasney, R. (2016). A framework for incorporating insurance in critical infrastructure cyber risk strategies. *International Journal of Critical Infrastructure Protection*, 14:43–57.

Zeller, G. and Scherer, M. (2021). A comprehensive model for cyber risk based on marked point processes and its application to insurance. *European Actuarial Journal*, pages 1–53.

Zheng, Y., Wang, J., Li, X., Yu, C., Kodaka, K., and Li, K. (2014). Driving risk assessment using cluster analysis based on naturalistic driving data. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2584–2589. IEEE.

Appendix A

Geographical Clusters

This table lists the a set of five clusters that is generated in Section 2.3.3, with data points distribution information.

Cluster Label	Number of Observations	States
1	3024	CT, DE, DC, IL, IA, ME, MD, MA, MI, NE, NH, NJ, NY, OH, PA, RI, SD, VT, WI, WY
2	531	AK, MN, MT, ND, OR, WA
3	1642	AL, AR, FL, GA, ID, LA, MS, OK, SC, TX
4	1093	IN, KS, KY, MO, NC, TN, VA, WV
5	1805	AZ, CA, CO, HI, NV, NM, UT

Appendix B

Variance-covariance Matrix

The posterior estimation of variance-covariance matrix $\mathbf{\Sigma}$:

$$\widehat{\mathbf{\Sigma}} = (\hat{\sigma}_{ij}) = \begin{pmatrix} 0.0893 & 0.0513 & 0.0763 & 0.0041 & -0.0002 & -0.0061 \\ 0.0513 & 0.1005 & 0.0641 & -0.0126 & -0.0097 & -0.0072 \\ 0.0763 & 0.0641 & 0.1166 & 0.0133 & 0.0066 & -0.0084 \\ 0.0041 & -0.0126 & 0.0133 & 0.0421 & 0.0115 & -0.0008 \\ -0.0002 & -0.0097 & 0.0066 & 0.0115 & 0.0199 & -0.0003 \\ -0.0061 & -0.0072 & -0.0084 & -0.0008 & -0.0003 & 0.0008 \end{pmatrix}$$

where $\hat{\sigma}_{ij}$ is the mean of posterior distribution of σ_{ij}.

Appendix C

Geographical Regions

The United States Census Bureau, divides the United States into four regions: the Northeast, the Midwest, the South, and the West. This classification rule is used in our frequency modeling and loss aggregation in order to reduce the number of scattered geographical locations.

Region	Division	States
Northeast	New England	CT, ME, MA, NH, RI, VT
	Middle Atlantic	NJ, NY, PA
Midwest	East North Central	IL, IN, MI, OH, WI
	West North Central	IA, KS, MN, MO, NE, ND, SD
South	South Atlantic	DE, FL, GA, MD, NC, SC, VA DC, WV
	East South Central	AL, KY, MS, TN
	West South Central	AR, LA, OK, TX
West	Mountain	AZ, CO, ID, MT, NV, NM, UT, WY
	Pacific	Alaska, CA, HI, OR, WA