Joint Model Prediction for Individual-Level Loss Reserving and a Framework to Improve Ratemaking in Non-Life Insurance

by

A. Nii-Armah Okine

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Business)

at the

UNIVERSITY OF WISCONSIN-MADISON

2020

Date of final oral examination: 06/26/2020

The dissertation is approved by the following members of the Final Oral Committee:

Peng Shi, Professor, Business
Edward W. Frees, Professor Emeritus, Business
Marjorie Rosenberg, Professor, Business
J. Tyler Leverty, Professor, Business
Zhengjun Zhang, Professor, Statistics

Acknowledgments

Many people contributed to my progress and the eventual completion of this degree. At the heart of this accomplishment are my advisors Prof. Peng Shi and Prof. Edward (Jed) Frees, who provided immense intellectual guidance and emotional support when I needed it. I want to express my profound gratitude to them for the support, trust, advice, and inspiration that motivated me throughout the process.

Secondly, I am sincerely grateful to my other committee members Prof. Margie Rosenberg, Prof. Tyler Leverty, and Prof. Zhengjun Zhang, for their insightful comments and valuable suggestions that helped improve the quality of this thesis.

Further, I am thankful to the School of Business and the Risk and Insurance Department for providing the conducive environment and facilities to help me complete my research work. The perfect blend of presentations from the statistics, economics, actuarial science, risk management, and insurance provided me with all the tools to see the bigger picture. I am also thankful to the Society of Actuaries for awarding me the Hickman Scholarship.

Fourth, I am thankful to all my fellow Ph.D. students from the Risk and Insurance department. Thanks for all those fun times we shared, and I appreciate all the thoughtful comments during our brainstorming and research meetings.

Finally, but most importantly, my reverent appreciation to my wife Abigail and son Jeremy for their patience, encouragement, unwavering support, and endless love that provided me with much-needed strength and motivation. I am also eternally thankful to my parents, Margaret Addoquaye and Thomas Okine, who have been my biggest fans.

Abstract

In non-life insurance, a loss reserve represents the insurer's best estimate of outstanding liabilities for losses that occurred on or before a valuation date. The accurate prediction of outstanding liabilities is key to setting reserves and calibrating insurance rates, which are two interconnected primary functions of actuaries. For instance, inadequate reserves could lead to deficient rates and thereby increase solvency risk. Also, excessive reserves could increase the cost of capital and regulatory scrutiny. Therefore, reserving accuracy is essential for insurers to meet regulatory requirements, remain solvent, and stay competitive.

The loss reserve prediction in non-life insurance is usually based on macro-level models that use aggregate loss data summarized in a run-off triangle. The main strengths of the macro-level models are that they are easy to implement and interpret. But, the limited ability to handle heterogeneity among triangle cells and changes to the business environment may lead to inaccurate predictions. Recently, micro-level reserving techniques have gained traction as they allow an analyst to use the information on the policy, the individual claim, and the development process to predict outstanding liabilities. Granular covariate information allows environmental changes to be captured naturally to improve reserve predictions.

In non-life insurance, the payment history can be predictive of the timing of a settlement for individual claims. Ignoring the association between the payment process and the settlement process could bias the prediction of outstanding payments. To address this issue, In this dissertation, I introduce into the literature of micro-level loss reserving

a joint modeling framework that incorporates longitudinal payments of a claim into the intensity process of the claim settlement. I discuss statistical inference and focus on the prediction aspects of the model. I demonstrate applications of the proposed model in the reserving practice and identify scenarios where the joint model outperforms macro-level reserving methods using simulated data. Moreover, I present a detailed empirical analysis using data from a property insurance provider. I fit the joint model to a training dataset and use the fitted model to predict the future development of open claims. The prediction results using out-of-sample data show that the joint model framework outperforms existing reserving models that ignore the payment-settlement association.

In pricing insurance contracts for non-life insurers, current methods often only consider the information on closed claims and ignore open claims. In case of a shift in the insurer's book risk profile, open claims could reflect the change in a timely manner compared to closed claims. This dissertation presents an intuitive ratemaking model by employing a marked Poisson process framework. The framework ensures that the multivariate risk analysis is done using the information on all reported claims and makes an adjustment for incurred but not reported claims based on the reporting delay distribution. Using data from a property insurance provider, I show that by determining rates based on current data, the proposed ratemaking framework leads to better alignment of premiums with claims experience. Among other things, accurate risk pricing suggests that all market participants, insurers, and customers, bear reasonable costs for risks assumed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Chapter Preview. In the loss reserving literature for non-life insurance, there are two main classes of reserving techniques: Macro-level models, which are sometimes called aggregate models, and micro-level models, also referred to as individual-level models. In this chapter, I review the strengths and limitations of these classes of reserving techniques. In claims management, actuarial analysts commonly encounter situations where settlement duration is positively associated with the size of payments for individual claims. This chapter also describes the features of insurance claims data that motivate the joint modeling of the payment and settlement processes to account for the association between them.

Section 1.1 discusses the concept of reserving. Section 1.2 then elaborates on macro-level and micro-level reserving techniques by discussing their strengths and limitations. Section 1.3 describes the property insurance claims dataset and its important characteristics that motivate the joint modeling framework. The chapter concludes by explaining the contribution of this dissertation to the loss reserving, joint model, and ratemaking literatures in Section 1.4.

1.1 The Concept of Loss Reserving

A loss reserve represents the insurer's best estimate of outstanding liabilities for losses that occurred on or before a valuation date. The process of estimating the outstanding liabilities is called loss reserving. In addition to loss reserves, non-life insurers make provisions for unearned premiums. The unearned premium reserve is the money insurers hold for premiums received, but for which coverage has yet to be provided hence a claim has not yet occurred. The sum of these two classes of reserves makes up the most substantial liability on a non-life insurer's balance sheet. The central focus of this dissertation, however, is on loss reserves. For more on unearned premium reserves, see Werner and Modlin (2016).

Substantial errors in predicting unpaid losses can have important business consequences. Inaccurate prediction of unpaid losses may lead to under-reserving (inadequate reserves) or over-reserving (excessive reserves), which influences the insurer's key financial metrics that further feeds into the decision making of management, investors, and regulators (Petroni, 1992). For example, over-reserving could increase the cost of capital. Further, as will be discussed in Chapter 5, in pricing insurance products through a process known as ratemaking, the accurate prediction of unpaid losses is of great value. For instance, under-reserving could lead to deficient insurance rates and thus increase solvency risk. Therefore, reserving accuracy is essential for insurers to meet regulatory requirements, remain solvent, and stay competitive. An expansive list of the importance of accurately estimating unpaid losses can be found in the first chapter of Friedland (2010).

At a valuation date, which is the date at which historical payment transactions are used to evaluate outstanding liabilities, a claim can be Incurred But Not Reported (IBNR), Reported But Not Settled (RBNS), or settled. Reserve estimates are then provided for IBNR, and RBNS claims to cover outstanding liabilities. The development process of a single non-life claim is illustrated in Figure 1.1. In the figure, the claim that occurred at time T_{occ} is notified to the insurer at time T_{rep}. After the claim is reported, it may take several payment transactions made at times T_1, T_2 and T_3 (sometimes just

one payment transaction) for the claim to be settled at time T_{set}. In this scenario, the claim is an IBNR claim at the valuation date τ_1, a RBNS claim at the valuation date τ_2, and a settled claim at the valuation date τ_3. The reporting delay, which is the difference between the reporting date and the occurrence date, can be affected by a number of factors. For example, usually big claims are easily noticeable hence are reported quicker than small claims. Also, the settlement delay, which is the difference between the settlement date and the reporting date, is affected by the time it takes to evaluate the whole size of the claim. As a result, the settlement delay is generally longer for claims with disputes which have to be settled in court.

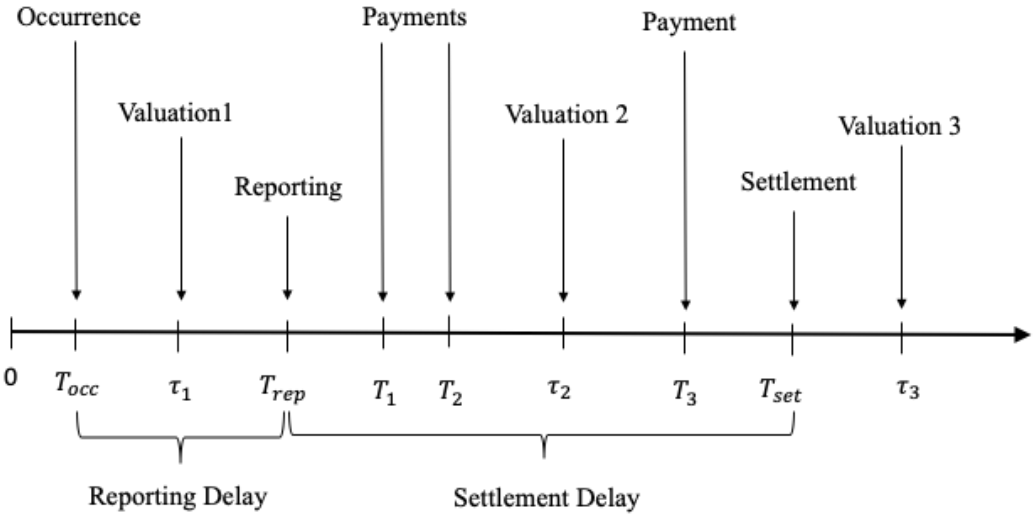

Figure 1.1: Timeline for the development of a non-life claim.

1.2 Literature

In the literature, there are two main classes of reserving techniques: Macro-level models, and micro-level models. This section reviews the strengths and limitations of these classes of reserving techniques.

1.2.1 Macro-Level Reserving Models

The macro-level models are based on aggregate losses data summarized in a run-off triangle. The run-off triangle, which is also called the loss development triangle, sums up losses using two time axes, claim occurrence period i and development period j, as shown in Figure 1.2. Here, the period can be years, half-years, quarters, months, etc., yet time is typically treated in a discrete manner. The data can be summarized as incremental payments $X_{i,j}$ defined as payments for claims in cell $\{i, j\}$ or cumulative payments defined as $C_{i,j} = \sum_{k=0}^{j} X_{i,k}$. The upper triangle is what we observe as the development of the losses is censored at the most recent accident period I. In the figure, it is assumed that the last development period is J. The observed data can be the paid losses, which are the amount paid out for claims or the reported losses, which refer to the cumulative paid losses plus the amount expected to be paid by claim adjusters. The lower triangle is made up of predicted amounts. An estimate for the total outstanding liabilities (combined RBNS and IBNR reserves) is obtained by subtracting the total paid (reported) losses from the predicted ultimate paid (reported) losses. Also, an estimate for only the RBNS reserve component can be obtained by organizing the run-off triangle using the reporting period and observation period instead of the occurrence period and development period.

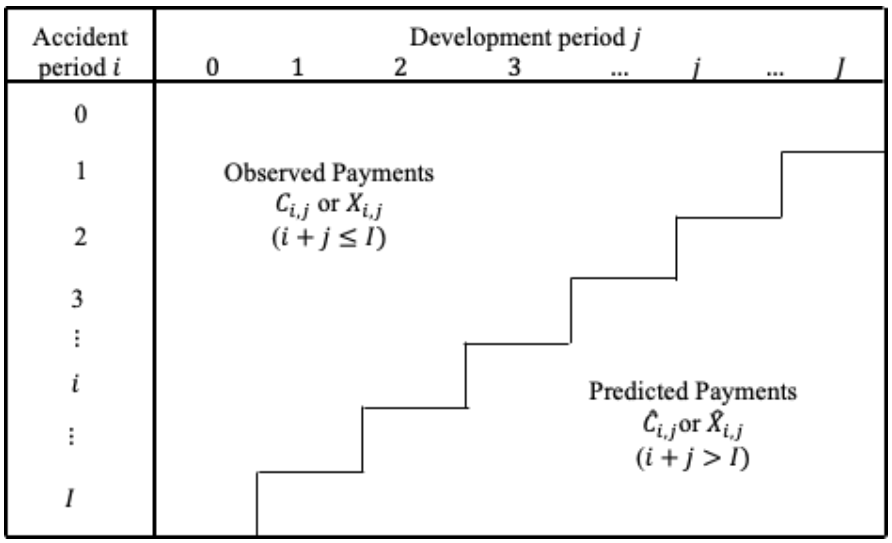

Figure 1.2: Loss run-off triangle.

1.2.1.1 Distribution-Free Models

The distribution-free macro-level models do not specify any distribution for the observed data in the upper triangle. The reserve is estimated using the chain-ladder (CL) method and its extensions. Before 1975, the actuarial literature explains the CL method as an algorithm for estimating loss reserves as a point estimate (Taylor, 2014). After then, the CL method has been motivated based on a stochastic model. The CL method links successive cumulative losses with appropriate link ratios, and Mack (1993) provides the first distribution-free stochastic model that underlies the CL algorithm. The distribution-free CL method assumes there exist $f_0, ..., f_{J-1} > 0$ such that for all $0 \leq i \leq I$ and all $1 \leq j \leq J$ we have :

$$E[C_{i,j}|C_{i,0}, ..., C_{i,j-1}] = E[C_{i,j}|C_{i,j-1}] = f_{j-1}C_{i,j-1}. \tag{1.1}$$

From (1.1), the conditional expected ultimate loss can be obtained. Thus, for all $1 \leq i \leq I$,

$$E[C_{i,J}|D_I] = E[C_{i,J}|C_{i,I-i}] = C_{i,I-i}f_{I-i}...f_{J-1}, \tag{1.2}$$

where D_I is the observed data in the upper triangle, and the factors f_j are called the link ratios, development factors, CL factors, or age-to-age factors. The predicted outstanding liabilities for each accident year would then be given by $E[C_{i,J}|D_I] - C_{i,I-i}$. The CL algorithm is characterized by (1.2) and involves the following steps: (1) calculating loss development factors; (2) selecting tail factors to bring losses to ultimate when development for the last development period is greater than 1; (3) calculating cumulative loss development factors (4) projecting the ultimate losses. To evaluate the variability in the reserve estimate from the CL method, Mack's model specified the first two moments of the underlying aggregate data and provided a formula for the prediction errors.

For accurate estimation of the development factors, the CL method relies on a stable

environment where there are no significant changes in the insurer's business that can affect loss reserving, for example, underwriting practices, claims processing, mix of products, and so forth. In an unstable environment or when data is not available because an insurer enters a new line of business, actuaries rely on expected loss techniques to estimate outstanding liabilities. An expected loss technique assumes that the actuary using their judgment can provide a better estimate of the total unpaid losses than using the observed losses experience.

The Bornhuetter-Ferguson (BF) technique (Bornhuetter and Ferguson, 1972), which is a combination of the CL technique and expected loss technique, allows the projection of ultimate losses based on actuaries' prior estimates. The BF method can be interpreted in a Bayesian framework, as experience matures more weight is given to actual losses, and judgments become less critical. The BF method becomes useful when there is instability in the proportion of ultimate losses paid in early development years, which could result in large reserve errors. Another deficiency in the CL model is that the ultimate loss entirely depends on the last observation on the diagonal (Wüthrich and Merz, 2008). If the last observation is an outlier, the outlier is projected to the ultimate loss. One way to deal with this is to make the diagonal observations more robust using the Cape-Cod method, which is also a combination of the CL technique and the expected loss technique. For more on the distribution-free macro-level models, see Friedland (2010).

To understand what reserving methods actuaries use to provide point estimates for reserves and model reserve variability, an ASTIN Working Party on Non-Life Reserving Practices conducted a survey on non-life reserving practices worldwide. ASTIN (2016) provides results from the survey, which show that the chain ladder is the most commonly used macro-level model for point estimates, followed by the BF method and the Cape-Cod method.

1.2.1.2 Parametric Stochastic Models

Another important family of macro-level models is the parametric stochastic models that assume a distribution for the aggregate data in a run-off triangle. Early stochastic models focused on reproducing reserve estimates as the chain-ladder technique; Hachemeister and Stanard (1975) shows that the Poisson model leads to chain-ladder estimates for reserves. Kremer (1982) specifies a log-normal distribution for the incremental losses. Mack (1991) proposes using gamma distribution for the distribution of incremental losses. Renshaw and Verrall (1998) implements an over-dispersed Poisson (ODP) model for incremental losses. Allowing for over-dispersion does not affect the estimation of parameters but increases standard errors. The ODP method cannot be applied if the column sum of incremental losses for any development period is negative. Verrall (2000) suggests a normal distribution whose support is not restricted to the positive real line for incremental losses with negative values. An overview of stochastic macro-level models can be found in England and Verrall (2002) and Wüthrich and Merz (2008).

In the literature, the mean square error of prediction (MSEP), also known as the prediction error, has been used as a measure of uncertainty in reserve estimates. The MSEP can be considered as the sum of two components, process variance which is as a result of variability in the data and estimation variance which comes from the uncertainty in the estimation of parameters of the reserving model (England and Verrall, 2002). In recent years, the full predictive distribution of reserve estimates with bootstrapping or Bayesian techniques has received considerable attention. With the predictive distribution, other information such as the skewness or risk measures of interest can be obtained. England and Verrall (1999) provides details on using bootstrapping to provide prediction errors for a GLM which reproduce reserve estimates of the chain-ladder technique. The results from the survey conducted by an ASTIN Working Party on non-life reserving practices worldwide also show the bootstrapping technique is the most used technique for reserve variability for macro-level models, followed by Mack's model (ASTIN, 2016).

Practicing actuaries find it relatively easy to implement the macro-level models as it

is less computationally challenging; hence, recent research has focused on addressing issues related to the CL method. Kuang et al. (2008) and Kuang et al. (2011) extended the chain-ladder model to deal with changes in the economic environment that affect policies for all accident years. Verrall et al. (2010), Martínez-Miranda et al. (2011), and Martínez-Miranda et al. (2012) propose a double chain-ladder method that uses the run-off triangle of paid losses and also the number of reported losses. Verrall et al. (2010) uses the compound Poisson framework that provides a clear split of estimates for IBNR and RBNS reserve. Further, it's expected that the additional information of the count triangle should lower the volatility of estimated reserves. Martínez-Miranda et al. (2011) based on the model developed in Verrall et al. (2010), constructs bootstrap estimates of the predictive distributions for the total reserve that splits into RBNS and the IBNR reserves. Martínez-Miranda et al. (2012) generalizes the model in Verrall et al. (2010) and Martínez-Miranda et al. (2011), to allow for loss inflation effect in the underwriting year direction. Martínez-Miranda et al. (2013) reformulates the triangular data as a histogram and proposes a continuous chain-ladder model through the use of a kernel smoother.

1.2.1.3 Limitations

There are three primary inherent limitations of macro-level models that cannot be addressed within the aggregate data framework. They are: (1) the limited ability to handle heterogeneity, (2) the limited ability to address environmental changes, and (3) issues of small sample size as a result of data aggregation.

An essential assumption of the macro-level models is that the losses aggregated in the run-off triangle are homogeneous. When practicing actuaries believe that the losses are heterogeneous, then they are often segmented by specific discrete characteristics and compiled into multiple triangles. This approach to addressing heterogeneity becomes problematic when the source of heterogeneity is not clear or is a continuous variable. Further, the reduction in the number of claims in each portfolio can lead to credibility

issues.

The limited ability to handle environmental changes is also one of the most significant drawbacks of the chain-ladder method and its extensions. For instance, Friedland (2010) examined the effects of environmental changes on reserve prediction and found that the chain-ladder type methods are appropriate only in a steady-state (stable environment). In the case of environmental changes, some of the commonly-used macro-models can generate a reserve estimate without material errors. To handle environmental changes, macro-level methods consider either expected loss techniques that allow actuaries to incorporate a priori reserve estimate or trending techniques that treat environmental change as a trend to adjust the development projections (Berquist and Sherman, 1977). However, highly dependent on actuaries' judgments, both techniques could lead to problematic reserve estimates (Thorne, 1978).

Moreover, another inherent issue with aggregation is that the observed data in the upper triangle is small. Concerns on fitting stochastic models on small sample size data in the run-off triangle were raised in Wright (1990), because it may lead to the poor choice of model. The small sample size could lead to a prediction error that could be disappointingly large (England and Verrall, 2002).

1.2.2 Micro-Level Reserving Models

In contrast to macro-level models that use aggregate data summarized in a run-off triangle to estimate outstanding liabilities, micro-level models use the information on the policy, the individual claim, and the development process to predict outstanding liabilities for each claim. As a result, the micro-level reserving techniques provide a Big Data approach to address the limitations of macro-level models. In recent years, following the general trend in analytics to look into detailed data, interests in micro-level techniques have spiked mostly because of their ability to leverage individual loss development to predict outstanding liabilities. Granular covariate information allows one to account for both claim and policy specific effects, and thus naturally captures the environmental changes.

Hence, reserve predictions from micro-level models are generally more accurate than those computed from aggregate data under non-homogeneous environmental conditions. For example, Guszcza and Lommele (2006) discusses the use of covariates to improve reserve estimates using claim-level data and illustrates the CL method's problems under a changing product mix with simulated data. Their results show that adding covariates to calculate the link factors captures the heterogeneous loss development patterns to improve predictions.

The use of individual-level data is not new to the insurance industry. For example, the individual-level approach is the norm in the life industry for pricing and reserving. Also, Parodi (2012) points out that non-life insurers use individual-level data for rate-making, but reserving is based on aggregate data though results from both pricing and reserving are needed to produce an overall model of the risk of an insurer. Therefore using micro-level modeling provides an opportunity for a consistent framework for both reserving and ratemaking in the non-life insurance industry.

1.2.2.1 Marked Poisson Process

The literature has made efforts to provide evidence for the advantages of loss reserving using individual claim data. The most studied method is the marked Poisson process framework introduced by Arjas (1989), Jewell (1989), Norberg (1993), and Norberg (1999). The marked Poisson process (MPP) represents events, such as claims or claim payments, as a collection of time points on a timeline with some additional features (called marks) measured at each point. This collection of randomly occurring time points can be represented using a specific type of stochastic process known as a counting process. The counting process of the events at the time points follows a non-homogeneous Poisson process, and the marks are also random variables that may have a time-dependent probability distribution. I provide a detailed description of the MPP framework below.

Let claims occur at times V_i satisfying $0 \leq V_1 \leq V_2 \leq \cdots$. The associated counting process $\{N(v), 0 \leq v\}$ is Poisson and records the cumulative number of claims that the

process generates. Specifically, $N(v) = \sum_{i=1}^{\infty} I(V_i \leq v)$ gives the number of claims in the interval $[0, v]$. Denote $H(v) = \{N(s) : 0 \leq s < v\}$ to be the history of the process at time v. For simplicity, it is assumed that two claims cannot co-occur for claims occurring on a continuous-time. The intensity function gives the instantaneous probability of a claim occurring at v, conditional on the process history. For example, in cases where there are no covariates, and only v determines the intensity, we have:

$$\rho(v|H(v)) = \lim_{\Delta v \downarrow 0} \frac{\Pr\{\Delta N(v) = 1|H(v)\}}{\Delta v} = \lim_{\Delta v \downarrow 0} \frac{\Pr\{\Delta N(v) = 1\}}{\Delta v} = \rho(v), \tag{1.3}$$

where $\rho(v)$ is a non-negative integrable function and $\Delta N(v)$ represents the number of claims in the short interval $[v, v + \Delta v)$. The Poisson process is seen in (1.3) as the intensity function is not affected by the process history at time v. When the intensity varies with v, the process is referred to as non-homogeneous; otherwise, it is said to be homogeneous.

For a Poisson process, the joint probability density for having n claims occur at times $V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n$ in a fixed interval $[0, \tau]$ is given by:

$$\Pr(N = n, V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n) = \exp\left(-\int_0^{\tau} \rho(v)dv\right) \prod_{i=1}^{n} \rho(v_i), \tag{1.4}$$

where $n \geq 0$. The Poisson process is the most common claim number process because it has attractive mathematical properties for insurance applications. For example, the finite-dimensional distribution of a Poisson process has a simple structure (Mikosch, 2009). The marginal probability of n claims can be shown to be given by:

$$\Pr(N = n) = \frac{\psi(\tau)^n e^{-\psi(\tau)}}{n!}, \quad n = 0, 1, \ldots, \tag{1.5}$$

which is a Poisson distribution with mean parameter $\psi(\tau) = \int_0^{\tau} \rho(v)dv$. In applications, $\rho(v)$ can be specified using a parametric model, and the common models include the

exponential, e.g. $\rho(v) = \exp(\alpha_0 + \alpha_1 t)$ and a piecewise-constant model.

We may include external covariates $x(v)$, which do not depend on the history of the counting process but may depend on time v, to account for the heterogeneities among the policyholders by specifying an intensity function of the form:

$$\rho(v|x^v) = \rho_0(v)\exp(x'(v)\beta), \tag{1.6}$$

where $x^v = \{x(s) : 0 \leq s \leq v\}$ is the covariate history. $\rho_0(v)$ is the baseline function that relates to policyholders for whom $x(v) = 0$ for all v, and β is a vector of regression coefficients for the covariates. Including internal covariates such as time since the most recent claim occurrence or the number of previous claim occurrences in (1.6) means the process is no longer Poisson but often referred to as modulated Poisson process (Cook and Lawless, 2007). When there is inter-policyholder variation even after conditioning on the external covariates, incorporation of unobserved random effects u_j can be considered for policyholders $j = 1, \ldots J$. Then the process $\{N_j(v), 0 \leq v\}$ is Poisson with intensity function:

$$\rho(v|x_j^v, u_j) = u_j\rho_0(v)\exp(x_j'(v)\beta), \tag{1.7}$$

where $u_1, \ldots u_J$ is taken to be i.i.d. with a distribution that satisfies $E(u_j) = 1$.

Formally, for a marked Poisson process in $[0, \tau]$, the probability density that n claims occur at times $V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n$, with marks $Z_1 = z_1, Z_2 = z_2, \ldots, Z_n = z_n$ is given by:

$$\Pr[N = n, (V_i, Z_i)) = (v_i, z_i), i = 1, 2, \ldots, n] = \exp\left(-\int_0^\tau \rho(v)dv\right)\left(\prod_{i=1}^n \rho(v_i)P_{Z|v_i}(z_i)\right), \tag{1.8}$$

where the claim occurrence counting process $N(v)$ is a Poisson process with intensity function $\rho(v)$. The distribution of the marks $P_{Z|v}$ is conditional on $\Delta N(v) = 1$.

The MPP framework allows for the modeling of the entire claim process, including occurrence, reporting, and development after reporting. Let $Z_i = (U_i, W_i)$, with U_i and W_i denoting the reporting delay and the claim development process after reporting, respectively. W_i includes transaction payments and settlement indicators. Then, $P_{Z|v} = P_{U|v} \times P_{W|v,u}$. The claims are random elements in the claim space $\mathcal{C} = [0, \infty) \times \mathcal{Z} = [0, \infty) \times [0, \infty) \times \mathcal{W}$ with intensity measure:

$$\rho(dv) \times P_{U|v}(du) \times P_{W|v,u}(dw), \quad (v, u, w) \in \mathcal{C}. \tag{1.9}$$

The MPP framework is sufficiently general to handle the different classes of claims involved in the loss reserving problem. With respect to the valuation time τ, claims can be decomposed into two subclasses, reported and IBNR claims, i.e. $\{C = C^{rep}, C^{ibnr}\}$. With reported claims, $\mathcal{C}^{rep} = ((v, u, w) \in \mathcal{C} | v + u \leq \tau)$, the full or partial development process is observed . But the development process is totally unobserved for IBNR claims, i.e. $\mathcal{C}^{ibnr} = ((v, u, w) \in \mathcal{C} | v \leq \tau, v + u > \tau)$. The disjoint subclasses decomposed from a marked Poisson process are also independent marked Poisson processes (Wüthrich and Merz, 2008). Then, the occurrence of reported claims follows an independent Poisson process with intensity function $\rho(v)F_{U|v}(\tau - v)$ and that of IBNR claims also follows an independent Poisson process with intensity function $\rho(v)(1 - F_{U|v}(\tau - v))$. More details on the MPP framework appear in Section 7.4.2.

Various special cases of this general setting have arisen in the actuarial literature. Jewell (1989) models the number of IBNR claims using a marked Poisson process, where the claim occurrence is assumed to follow a homogeneous Poisson process and the reporting delay treated as marks. In Arjas (1989), the development of losses is viewed as a marked Poisson process where the occurrence of transactions follow a non-homogeneous Poisson process, and transaction payment amounts are treated as marks. Norberg (1993; 1999) handle loss reserving in non-life insurance using the marked Poisson process, where

claims are assumed to occur by following a non-homogeneous Poisson process, and other stochastic characteristics about the claims are treated as the time-dependent marks. The set of all claims is divided into settled, reported-not-settled, incurred-not-reported, and covered-not-reported (corresponding to unearned premium reserve). It was asserted that claims from these subgroups will follow an independent marked Poisson process.

Moreover, Larsen (2007) specifies a discrete model that divides claims by occurrence year, reporting delay in years, and discrete characteristics. The claims in each subgroup are assumed to follow an independent marked Poisson process, and the likelihood function for each subgroup is specified. The paper illustrated the parameter estimation process and presented prediction results based on a small case study with data from a Marine insurance portfolio. Antonio and Plat (2014) provided the first detailed empirical study with data from a personal-line general liability insurance portfolio. The claim occurrence times are taken as the points and follow a Poisson process with non–homogeneous intensity. The marks are considered to be the reporting delay, transaction times, transaction types, and payment amounts. They provide a detailed routine to predict future loss development, and their results show that the micro-level model outperforms the results obtained with traditional loss reserving methods for aggregate data. Verrall and Wüthrich (2016) specifies a non-homogeneous Poisson process so that the model for the number IBNR can cope with trends and with seasonal patterns. They consider weekly periodic patterns and the reporting delay distribution is allowed to depend on the weekday of the occurrence of the claims. Their model was calibrated to a property and casualty insurance datasets, and their results show that the individual-level model performs better than the chain-ladder model.

1.2.2.2 Marked Cox Process

A Cox process, or doubly stochastic Poisson process, extends a Poisson process by modeling the intensity as a non-negative stochastic process. Avanzi et al. (2016) proposes a model to predict the number of IBNR claims by modeling the claim arrival process

along with its reporting delays as a marked Cox process to allow for overdispersion and serial dependency. The paper notes that even though the Poisson process increases the flexibility that accurately represents the nature of claim frequencies, the deterministic intensity function under the Poisson process does not allow for the serial dependency of claims counts. The shot noise intensity was used with a filtering algorithm to filter out the underlying intensity process and estimate the parameters accurately.

Badescu et al. (2016b) and Badescu et al. (2016a) proposed a marked Cox process to model the temporal dependence exhibited in the claim arrival process. They provide a generalized expectation-maximization (EM) algorithm which guarantees the efficiency of the estimators unlike the moment estimation methods widely used in estimating Cox processes. They show that the marked Cox process performs better than predictions from the ODP model.

1.2.2.3 Generalized Linear Models

Another family of research using individual-level data employs generalized linear models (GLMs) in conjunction with survival analysis to incorporate settlement time as a predictor for ultimate losses. Using the GLM enables the investigation and modeling of many features of the data responsible for the violation of CL conditions.

The Generalized Linear Models (GLMs) extend stochastic modeling from Gaussian distribution to distributions from the exponential family. Various outcomes can be modeled using the GLM framework, e.g., binary, counts and skewed outcomes. To define the GLM, let $Y_1, \ldots, Y_n$ be independent random variables with a distribution from the exponential family; the general formulation is given by:

$$f(y_i) = \exp\left(\frac{y_i\theta_i - \psi(\theta_i)}{\phi} + S(y_i, \phi)\right), \tag{1.10}$$

where $\psi(\cdot)$ and $S(\cdot)$ are known functions, and θ_i and ϕ are the natural and dispersion

parameters respectively. It can be shown that:

$$\mu_i = E[Y_i] = \psi'(\theta_i) \quad \text{and} \quad Var[Y_i] = \phi\psi''(\theta_i) = \phi V(\mu_i), \tag{1.11}$$

where derivatives are with respect to θ and $V(\cdot)$ is the variance function. With GLMs, a transformation of the mean is modeled as a linear combination of covariates via a link function $g(\cdot)$. This gives:

$$\eta_i = g(\mu_i) = \mathbf{x}_i'\beta \tag{1.12}$$

Here, $\mathbf{x}_i$ are the vectors of covariates, and β is the regression coefficients to be estimated.

For reserving applications, Y_i may denote the ultimate loss for claim i and x_i is the characteristics of claim i, which include the settlement times T_i. Then the estimated model may be used to predict the ultimate loss for currently open claims. Because the data used to fit the model only contains closed claims, this approach does not utilize information from the longitudinal payment trajectory and makes it impossible to model each claim's dynamic development.

An estimate of the settlement times is required to predict the ultimate loss for open claims using the GLM model. A natural approach to estimating the settlement time is survival analysis, and a popular framework is the proportional hazards model (Cox, 1972). These models assume covariates have a multiplicative effect on the hazard function of settlement time T_i given by:

$$h_i(t|w_i) = h_0(t)\exp\{\gamma' w_i\}, \tag{1.13}$$

where w_i is a vector covariate assumed to be associated with the hazard of each claim with a corresponding vector of regression coefficients γ. The baseline hazard function $h_0(t)$ relates to the hazard function of a claim that has $\gamma' w_i = 0$. The baseline hazard function may take parametric and non-parametric forms. For more on the analysis of event time data, see Rizopoulos (2012) and Elashoff et al. (2016).

Several forms of individual claim model based on the GLM approach for loss reserving has been proposed in the loss reserving literature. Taylor and Campbell (2002) proposes GLMs and survival analysis for predicting case estimates with specific examples using data from a worker's compensation insurance. Taylor and McGuire (2004) modeled the total amount paid per finalized claim with GLMs. Taylor et al. (2008) models individual claims with GLMs and categorized covariates into three groups: covariates like the type of vehicle that do not change, covariates like development periods that predictably vary over time, and unpredictable dynamic covariates like health condition of a claimant that unpredictably change over time. Inclusion of the unpredictable dynamic covariates will require another model to predict these uncertain quantities. The paper employed the proportional hazards model for estimating the settlement times. Their results showed that the individual models exhibited higher predictive efficiency than aggregate models.

1.2.2.4 Machine Learning Algorithms

The individual-level models discussed so far assume a fixed structural form. Recently, another stream of research for individual-level reserving focuses on using machine learning algorithms. According to Wüthrich (2018a), machine learning algorithms have become popular because they are highly flexible and can deal with any structured and unstructured claim information. Machine learning algorithms may provide better prediction accuracies than linear models because they take advantage of non-linear forms that provide a data-driven approach. Commonly used statistical learning methods that exploit non-linear relationship and interaction among features include decision trees, bagging, random forest, boosting, neural networks, and support vector machines. See Hastie et al. (2009) for more on these methods.

Wüthrich (2018a) paper illustrates the use of regression trees for individual-level reserving and focuses on modeling the number of payments in a discrete-time setting by incorporating both dynamic and static covariates. Results using the regression tree were compared to a 'homogeneous' estimate, which does not consider any feature information,

and the results show that the feature consideration is relevant for individual losses prediction. Baudry and Robert (2019) proposed a flexible approach for estimating the individual IBNR and RBNS losses separately using the ExtraTrees machine learning algorithm. De Felice and Moriconi (2019) presents a compound frequency-severity model for reserve prediction based on regression trees, and Duval and Pigeon (2019) proposes models for non-life individual-level loss reserving using the gradient boosting algorithm.

Another machine learning algorithm that has been studied in the loss reserving literature is the neural networks. Wüthrich (2018b) extends the CL method using neural networks by incorporating heterogeneous individual claims feature information in modeling the CL factors. This extended model allows for capturing environmental changes. ASTIN (2017) summarizes the work accomplished by a team of people that met within the scope of an ASTIN working party named Individual Claim Development with Machine Learning. The goal of the working party was to research the field of machine learning in connection with reserving as traditional actuarial work. Their work focused on using artificial neural networks (ANN) and was implemented in a cascading triangular way. The prediction results were compared with results achieved by classical reserving methods. Their results show that, if a line of business is not homogeneous, ANN performs better than the chain-ladder technique. Poon (2019) also presents a granular machine learning model framework to predict loss development using neural networks.

1.3 Motivation

1.3.1 Accurate Estimation of Reserves

As discussed earlier, though the CL technique is easy to implement in practice, it comes with a risk of inaccurate predictions. To show that the aggregation of data leads to information loss that can impact outstanding liabilities predictions, I use data from the Wisconsin Local Government Property Insurance Fund (LGPIF), which provides property coverage for local government units such as counties, cities, towns, villages,

school districts, library boards, etc. More on the background and description of the LGPIF can be found in Section 4.1. Here, the data used in the CL model calibration is restricted to the building and contents coverage, and to claims that have occurred and were reported between January 1, 2006, and December 31, 2009. It is assumed that at the end of this time window on December 31, 2009, (which is considered to be the valuation date), the Fund sets capital aside to cover future payments related to the reported claims. The actual development of the RBNS and IBNR claims is contained in the validation sample, which is from January 1, 2010, to December 31, 2013.

Based on data in the training dataset, Table 1.1 summarizes the distribution of the continuous covariates and the two outcomes of interest, i.e., the ultimate loss and the settlement duration (closed date minus the reported date plus one day). The significant associations between the continuous covariates and the outcomes of interest, as shown by the Spearman correlations (ρ_S), suggest that they will be useful for predicting outstanding payments. The Spearman correlation (ρ_S) between the ultimate losses and settlement times is 0.49. The positive association indicates that bigger claims take a longer time to settle; more on the association between the payment size and settlement is discussed in Section 1.3.2.

Table 1.1: Summary statistics for closed claims.

	Min.	Median	Mean	Max.	Ultimate Loss (ρ_S)	Settlement Time (ρ_S)
Ultimate Loss	25	2,203	14,133	2,633,822	-	0.49
Settlement Time (Days)	1	38	66	861	0.49	-
Deductible	500	1,000	12,297	100,000	-0.28	-0.21
Initial Estimate	30	2,500	9,545	1,000,000	0.93	0.51
Reporting Delay (Days)	0	28	66	864	-0.29	-0.55

The Kaplan-Meier estimate of survival probabilities of the School, City, and miscellaneous (Misc) entities is provided in Figure 1.3. The miscellaneous entity type includes fire stations. The survival probabilities at time t represent the probability that claims are not settled at that time. It seems that the claims from Schools have slightly higher survival probabilities than the other entity types.

The left panel of Figure 1.4 shows the number of claims that occurred in each quarter

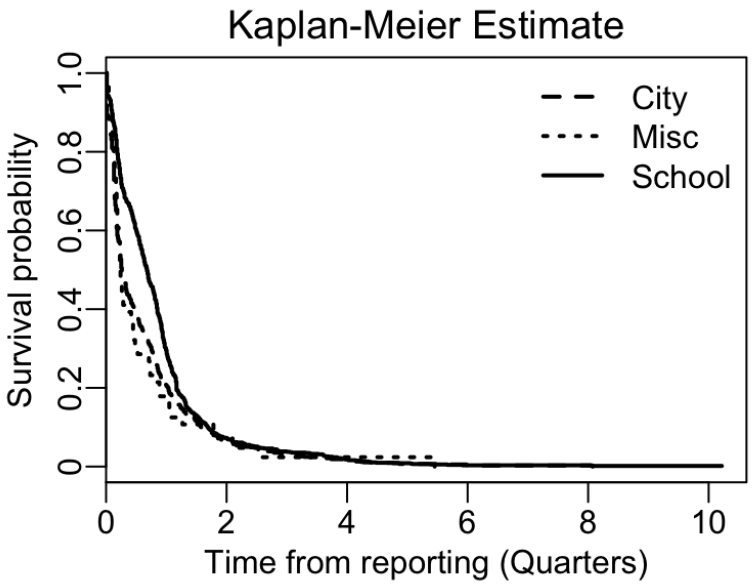

Figure 1.3: Survival probability plot by Entity type (Kaplan-Meier estimate).

from January 2006 to December 2009. Similar seasonal fluctuations are observed over each year, with the lowest occurrence in the winter season. The right panel of Figure 1.4 shows the distribution of the reporting delays in quarters. Approximately 75% of the claims are reported within the first quarter of the accident occurrence, but the distribution appears to be highly skewed to the right. Note that a reporting delay of zero corresponds with reporting on the day of occurrence. Also, the low number of reported claims in the year 2009, as seen in the left panel of Figure 1.4, is due to IBNR claims. Other supporting figures and tables are provided in Appendix 7.1. Figure 7.1 shows claims in the LGPIF data by region. It can be seen that there are some variabilities in the number of claims by region.

To use the CL approach to predict the total liabilities, which is the sum RBNS and IBNR reserve estimates, the individual paid losses data is aggregated in a run-off triangle using two time axes, claim occurrence quarter, and development quarter. Here, I employ the Mack CL model (Mack, 1993), and the analysis was performed in R following the

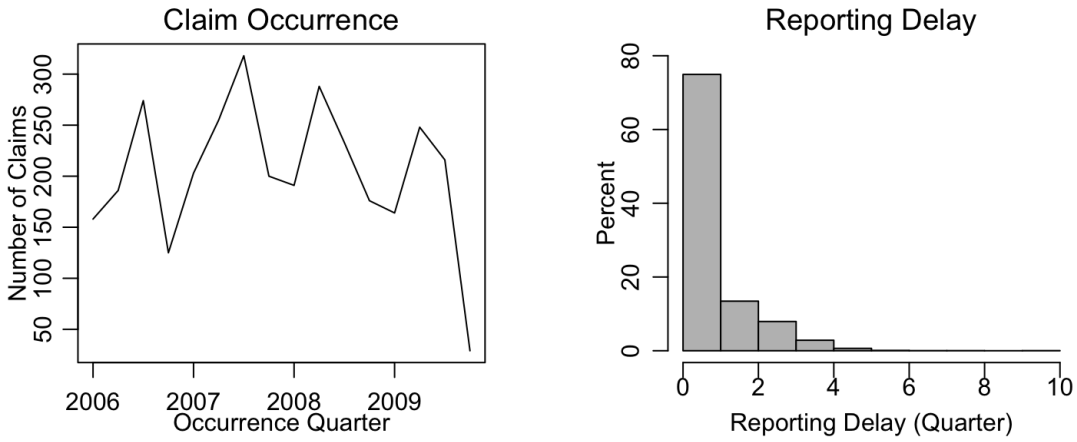

Figure 1.4: Left Panel: Number of claims occurred in each quarter from January 2006 to December 2009. Right Panel: Reporting Delay.

`ChainLadder` package (Carrato et al., 2020). Furthermore, to use the CL approach for the estimate of RBNS reserve, I employ a modified version of the run-off triangle where the individual losses data is aggregated using the reporting quarter and observation quarter instead of the occurrence quarter and development quarter. Then projections made from these developments factors give us RBNS reserve estimates.

Table 1.2: Prediction performance using the CL approach.

	RBNS		Total Liabilities	
	Estimate	Error %	Estimate	Error %
True Reserve	4,511,490		6,298,298	
Chain-Ladder Error	1,261,332	27.96	3,064,504	48.66

Table 1.2 presents the prediction performance using the CL approach for both RBNS and total liabilities. The percentage error for the estimate of the RBNS reserve liabilities is 27.96%, and that of the total unpaid liabilities is 48.66%. The results show the chain-ladder method did not perform well in estimating the total outstanding liabilities and RBNS reserve. One possible reason is the limitations of the aggregated data. As seen, the LGPIF data is heterogeneous and necessary information to control for the loss development process, i.e., covariate information about the policy, policyholder, and the claim is available only on the individual claim level. Therefore, to improve on the

predictions from the chain-ladder algorithm, an individual-level reserving model may be useful.

1.3.2 Association of Payment History and Settlement

In claims management, it is common that small claims are settled faster than large claims, because large and complicated claims naturally require experienced adjusters, demand special expertise, involve multiple interested parties, and are more likely to be litigated. As a result, the duration of settlement and size of payments for individual claims are often positively correlated. To illustrate, Figure 1.5 shows such a relationship using data from the LGPIF. I plot the distribution of ultimate payments against the settlement time in quarters (days/366 $\times$ 4) for a random sample of claims from the building and contents coverage. The solid line in the right panel is the fit of the loess scatterplot smoother. Both plots suggest a strong positive relation between ultimate payment and settlement time, i.e., it takes longer time periods to close larger claims. The insight provided from the payment-settlement association plot in Figure 1.5 suggests that we can do a better job in reserve prediction by incorporating the payment-settlement association in the prediction process as the development of payment may yield early indications of an impending settlement.

The payment-settlement association has important implications for the loss reserving practice. In loss reserving, actuaries predict the outstanding liabilities based on the claim history that is only observed up to a valuation date. When the settlement time and claim size are correlated and not accounted for, the historical claims that actuaries use for model building will not be representative of future payments, because large claims with longer settlement times will be more likely to be censored (not settled) by the valuation date, a type of selection bias. Specifically, when larger claims take more time to settle, outstanding payments will be underestimated if the selection bias in the sampling procedure is not accounted for. Similarly, one would expect overestimation of future payments if the claim size and settlement time were negatively correlated.

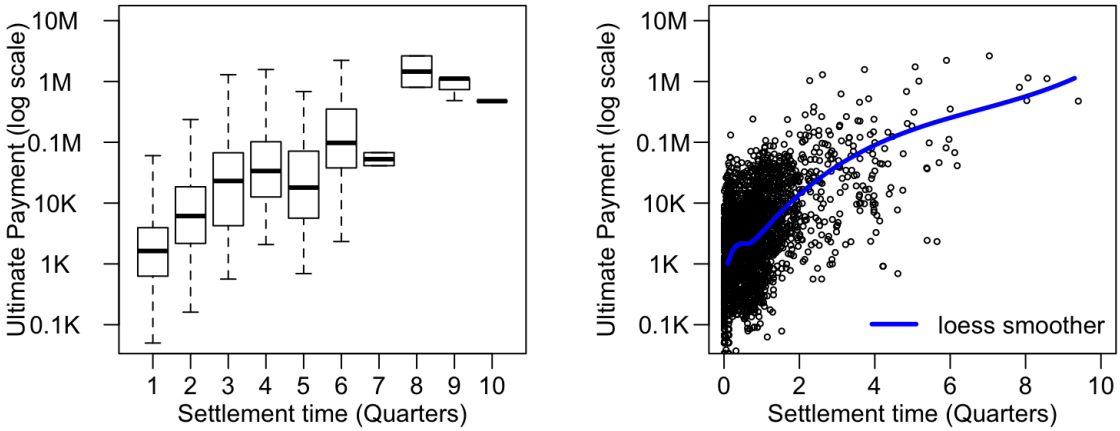

Figure 1.5: Distribution of ultimate payments by settlement time using data from a property insurer.

Further, the payment-settlement association means that payment history may help predict settlement time, which in turn feeds back into the prediction of unpaid losses. Then the relation between the two processes allows for the dynamic prediction of outstanding liabilities. The prediction is dynamic because, at a future date, when more information becomes available, the settlement time and ultimate payment predictions can be updated. The dynamic prediction entails two steps; the first step involves using the payment history and the fact that the claims are open at the valuation date to update the settlement time's prediction. For the second step, with the updated settlement time and the payment history, more accurate predictions for the ultimate payment can be made.

1.4 Contribution to Literature

1.4.1 Capture Payment-Settlement Association

The main goal of this dissertation is to establish a micro-level loss reserving method that leverages claim level granular information while accounting for the payment-settlement association, and thus improves accuracy in loss prediction. In doing so, I employ a joint modeling framework developed in the statistical literature for longitudinal outcomes and time-to-event data. The joint model for reserving purposes consists of two submodels, a longitudinal submodel that governs the payment process for a given claim and a survival submodel that concerns the settlement process of the claim. The two components are joined via shared latent variables. The joint model has a natural interpretation in the reserving context, where the historical payments affects the instantaneous settlement probability and the settlement intensity determines whether there are further payments. The joint model framework provides a novel solution to the sample selection issue that is due to the association between the size of claims and time of settlement.

1.4.2 Improvement of Existing Reserving Models

The joint model framework improves the accuracy in loss prediction compared to macro-level models by leveraging claim level granular information to control for heterogeneity and environmental changes. In this dissertation, I identify several scenarios of unstable environments where the standard chain-ladder method fails, while the joint model demonstrates superior performance. Similarly, the joint model framework offers an improvement over the existing individual-level reserving models by explicitly accounting for the payment-settlement association, hence addressing the issue of selection bias.

1.4.3 RBNS Prediction Using the Joint Model Framework

Properties of the joint model have been well-developed in the biomedical literature in clinical studies (Ibrahim et al., 2010) and non-clinical studies (Liu, 2009). But, the existing literature has primarily focused on the estimation aspect of inference. For statistical inference of the joint model, I discuss both estimation and prediction with the focus on the latter. The properties of estimators are investigated using extensive simulation studies in Chapter 3. Because of the predictive nature of loss reserving, I investigate the joint model's predictive performance using extensive simulation studies and a detailed empirical study, which enriches the existing statistical literature.

1.4.4 Detailed Empirical Analysis

In this dissertation, I present a detailed empirical analysis of the joint model framework using data from a property insurance provider with the focus on RBNS reserve prediction in Chapter 4. I fit the joint model to a training dataset, and the association between the payment history and settlement time is captured, which helps to accurately predict the settlement time and the ultimate amount of unsettled losses. The RBNS prediction performance of the joint model is compared to existing reserving models using an out–of–sample data. Further, because of the time dimension involved with the RBNS reserve prediction, the traditional cross-validation techniques cannot be used to evaluate the robustness of the prediction results. Thus, I introduce a novel form of cross-validation for longitudinal data that I call double cross-validation.

RBNS reserves are essential to actuaries. Generally, for occurrence-based policies, where coverage is triggered by date the claim occurred, RBNS reserves make up a more significant portion of the total reserves for fast-reporting lines like auto and homeowners, which is true for this dataset. Also, for claims-made policies, where coverage is triggered by date claim is reported, insurers are only responsible for the claims reported during the year, which is known by the end of the year. Hence, IBNR reserves are usually not

required for claims-made policies. Though the analysis focuses on RBNS reserves, in Section 4.3.4, I provide a discussion on how to apply the joint model framework for IBNR reserves.

1.4.5 Bridging the Gap Between Reserving and Ratemaking

In Chapter 5 of this dissertation, I provide a framework to improve insurance pricing using the marked Poisson process (MPP) framework. As described in Section 1.2.2, the MPP framework has primarily been used for individual-level reserving. Therefore, by implementing it in a ratemaking exercise, I bridge the gap between the two processes. Chapter 5 contributes to the ratemaking literature in two main ways. First, the proposed framework ensures that the multivariate risk analysis is done using the information on both open and closed claims in a more efficient way leading to better alignment of premiums to claims experiences. Second, by automatically accounting for the expected cost relating to both RBNS and IBNR claims without using a separate reserving model, the proposed framework makes the ratemaking process complete and balanced for individual risks.

Plan for Remaining Chapters: Chapter 2 introduces the joint model framework for individual-level loss reserving. In Chapter 3, a simulation study is conducted to investigate the estimation and prediction performance of the joint model framework. I find that the joint model displays superior outstanding liabilities prediction performance over the standard chain-ladder method under several scenarios of unstable environments. Further, the results show that micro-level models that ignore the payment-settlement association produce biased parameter estimates and inaccurate outstanding liabilities predictions, even under stable environmental conditions. Chapter 4 implements the joint model using claims data from a property insurance provider to evaluate its performance using real-world data. I focus on RBNS reserve prediction. The results show that accounting for the payment-settlement association leads to better prediction accuracy and lower reserve uncertainty than models that ignore it. Chapter 5 employs the

marked Poisson process framework to improve insurance pricing. The framework allows the optimal use of information on all reported claims hence promotes equity in the ratemaking process. Chapter 6 provides concluding remarks.

Chapter 2

Joint Model for Claim Payment and Settlement

Chapter Preview. In this chapter, the joint model framework is introduced to the loss reserving problem, focusing on a subset of selection models called shared-parameter models. In shared-parameter models, a latent random effects $\mathbf{b}_i$ is used to capture the association between the longitudinal and the time-to-event outcomes (Rizopoulos, 2012).

Section 2.1 reviews the literature on joint models for longitudinal and time-to-event data. Section 2.2 introduces the joint modeling framework for individual-level loss reserving. Sections 2.3 and 2.4 describe the specifications for the longitudinal submodel of claim payments and the survival submodel of claim settlement, respectively. To conclude, Section 2.5 discusses estimation and prediction for the joint model.

2.1 Background on Joint Longitudinal-Survival Models

The existing micro-level reserving methods do not explicitly capture the dependence between the payment history and settlement process. I further extend the literature by introducing the joint longitudinal-survival model (JM) framework to allow for such an association.

The joint model has been proposed in the medical statistics literature for modeling longitudinal and survival outcomes when the two components are correlated (Elashoff et al., 2016). DeGruttola and Tu (1994) model the progression of CD4-lymphocyte count and the relationship between different features of this progression and survival time using joint models. The paper considers random-effects models for repeated measures of CD4-lymphocyte count among AIDS patients receiving treatment with zidovudine (ZDV). Because many such patients do not survive throughout the study period, and the probability of death is related to the CD4-lymphocyte count, models for progression must take into account the missing CD4-lymphocyte counts caused by attrition. Their results show that the joint modeling framework provides efficient and unbiased estimates in contrast to biased estimate obtained from a two-stage model (Tsiatis et al., 1995). Furthermore, joint models have been applied in studies involving cancer patients. Chi and Ibrahim (2006) proposed a joint model that was motivated by a clinical trial conducted by the International Breast Cancer Study Group (IBCSG) to capture the unique features in the data set.

In a non-clinical setting, Liu (2009) proposes a joint random-effects model of longitudinal medical cost data and survival, taking into account the semi-continuous nature of medical costs. As a result of the semi-continuous nature of the medical cost, the paper employs a two-part model for the longitudinal medical costs, and the random effects from the two-part model are incorporated into the survival hazard model. The assumption here is that the mortality risk is correlated with both the frequency of hospital visits and

the amount of cost for medical treatment. That is, a sicker patient who is at a greater risk of death tends to seek medical treatment more often (thus have higher odds of positive monthly cost) and receive more intensive care when treated, resulting in a higher amount of monthly expenditure. Tsiatis and Davidian (2004), Yu et al. (2004), and Verbeke et al. (2010) give excellent overviews of joint models. Besides, Rizopoulos (2010) and Rizopoulos (2016) develop R packages for joint models.

Two general frameworks have received extensive attention, the pattern mixture model, and the selection model (Little, 2008). Let T_i be the time-to-event, and $\mathbf{Y}_i$ be a vector of longitudinal measurements for subject i; these two frameworks differ in the way the joint distribution $f_{\mathbf{Y}_i,T_i}(\mathbf{y}_i, t_i)$ is factorized and are discussed further in the following subsections.

2.1.1 Pattern Mixture model

For pattern mixture models, the joint distribution of the longitudinal and survival outcomes is specified using the marginal distribution of time-to-event outcome and the conditional distribution of longitudinal outcomes given the time-to-event outcome, i.e. $f_{\mathbf{Y}_i,T_i}(\mathbf{y}_i, t_i) = f(\mathbf{y}_i|t_i)f(t_i)$. Thus, the marginal distribution of the measurements can be viewed as a mixture of distributions (Little, 1993). As discussed earlier, the GLM approach employed in Taylor et al. (2008) follows a pattern mixture model.

2.1.2 Selection Model

In contrast, the joint distribution in selection models is specified using models for the marginal distribution of longitudinal outcomes and the conditional distribution of time-to-event outcome given longitudinal outcomes i.e., $f_{\mathbf{Y}_i,T_i}(\mathbf{y}_i, t_i) = f(\mathbf{y}_i)f(t_i|\mathbf{y}_i)$. The name for the selection models comes from Heckman (1976) from the economic literature. The conditional distribution $f(t_i|\mathbf{y}_i)$ is viewed as a probability that the subject self-selects to either continue or drop-out of study. Diggle and Kenward (1994)

was first to apply selection models to non-random drop-out in longitudinal studies by allowing the drop-out probabilities to depend on the history of measurement process up to the drop-out time.

2.2 General Framework

For the loss reserving application of the joint modeling framework, the sequence of payments from a reported claim forms the longitudinal outcomes, and the settlement time of the claim is the time-to-event outcome of interest. The development of claim payments may yield early indications of impending settlement, which introduces associations between the longitudinal and survival outcomes.

In this study, the reporting time is set as the time origin for a claim. For the ith claim $(i = 1, \ldots, N)$, I denote T_i^* and c_i as the settlement time and valuation time, respectively. Assuming c_i is independent of T_i^*, define $T_i = \min(T_i^*, c_i)$ and $\Delta_i = I(T_i^* < c_i)$, where $I(A) = 1$ when A is true and $I(A) = 0$ otherwise. The pair (T_i, Δ_i) makes up the observable time-to-settlement outcomes for claim i, where Δ_i indicates whether the claim has been closed by the valuation time; if so, T_i indicates the settlement time. Let $\{Y_i(t) : 0 \leq t \leq T_i^*\}$ be the payment process, and $\mathbf{Y}_i^* = \{Y_{it}, t \in \tau_i^*\}$ be the vector of the complete cumulative payments for claim i with n_i^* payments at times $\tau_i^* = \{t_{ij}; j = 1, \ldots, n_i^*\}$. Assume there are n_i payments by the time of valuation, define $\tau_i = \{t_{ij}; j = 1, \ldots, n_i\}$ as the observable payment times and denote $\mathbf{Y}_i = \{Y_{it}, t \in \tau_i\}$ the vector of cumulative payments at observed time of payments. Further denote $\mathbf{Y}_i^+ = \{Y_{it}, t \in \tau_i^+\}$ the vector of cumulative payments at future times $\tau_i^+ = \{t_{ij}; j = n_i + 1, \ldots, n_i^*\}$ after the valuation time. Let $\mathbf{b}_i$ be the vector of random effects that account for the claim-specific unobserved heterogeneity. Then the joint distribution $f_{\mathbf{Y}_i^*, T_i^*}(\mathbf{y}_i^*, t_i^*)$ is given by:

$$f_{\mathbf{Y}_i^*, T_i^*}(\mathbf{y}_i^*, t_i^*) = \int f(\mathbf{y}_i^* | \mathbf{b}_i) f(t_i^* | \mathbf{b}_i) dF(\mathbf{b}_i). \tag{2.1}$$

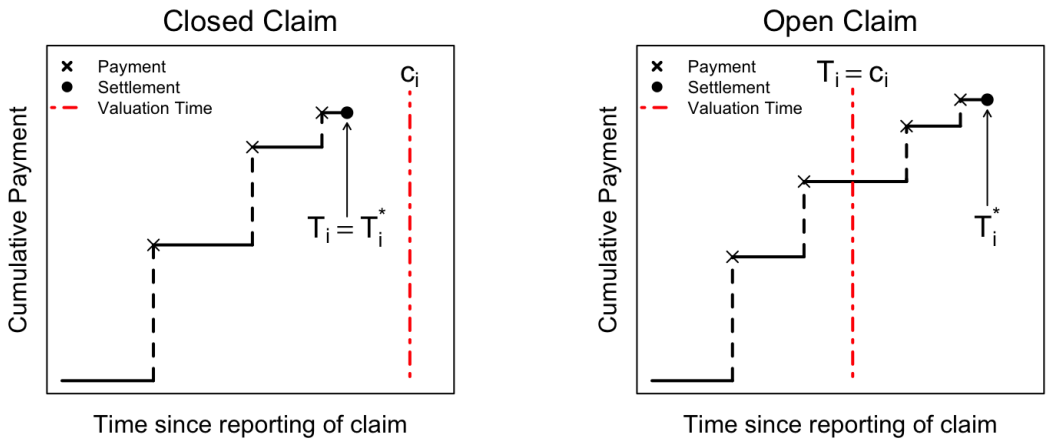

Figure 2.1: Graphical illustration of the cumulative payment process from the time of reporting to settlement.

Figure 2.1 provides a graphical illustration of the cumulative payment process that experiences jumps at the time of each payment from the time of reporting to settlement. The left panel presents a closed claim where the entire development process of the claim is observed before the valuation time, i.e. ($\Delta_i = 1$, $n_i = n_i^*$). The right panel provides an example of an open claim where only a part of the development process of the claim is observed at the valuation time, i.e. ($\Delta_i = 0$, $n_i \leq n_i^*$).

2.3 Longitudinal Submodel of Claim Payments

The cumulative payments Y_{it} is specified using generalized linear mixed effect models (see, for instance, Frees (2004) and Molenberghs and Verbeke (2006) for details). Conditional on the random effects $\mathbf{b}_i$, the cumulative payment Y_{it} is assumed to be from the exponential family

$$f(y_{it}|\mathbf{b}_i) = \exp\left(\frac{y_{it}\theta_{it} - \psi(\theta_{it})}{\phi} + S(y_{it}, \phi)\right), \tag{2.2}$$

where $\psi(\cdot)$ and $S(\cdot)$ are known functions, and θ_{it} and ϕ are the natural and dispersion parameters respectively. It can be shown that:

$$\mu_{it} = E[Y_{it}|\mathbf{b}_i] = \psi'(\theta_{it}) \quad \text{and} \quad Var[Y_{it}|\mathbf{b}_i] = \phi\psi''(\theta_{it}) = \phi V(\mu_{it}), \tag{2.3}$$

where $V(\cdot)$ is the variance function. The conditional mean is specified as a linear combination of covariates via a link function $g(\cdot)$, i.e.

$$\eta_{it} = g(\mu_{it}) = \mathbf{x}'_{it}\beta + \mathbf{z}'_{it}\mathbf{b}_i. \tag{2.4}$$

Here, $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are the vectors of covariates in the fixed and random effects, respectively, and β is the regression coefficients to be estimated. In this model, it is assumed Y_{it} are independent conditional on random effects $\mathbf{b}_i$. In addition, $\mathbf{b}_i$ are independent and follow a multivariate normal distribution, i.e., $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$. $\mathbf{D}$ is the covariance matrix for the random effects with unknown parameters ν.

2.4 Survival Submodel of Claim Settlement

The time-to-settlement outcome of a claim is modeled using a proportional hazards model. The hazard function of T_i^* is specified as:

$$h_i(t) = h_0(t) \exp\{\gamma' \mathbf{w}_{it} + \alpha\eta_{it}\}, \tag{2.5}$$

where $h_0(t)$ is the baseline hazard, $\mathbf{w}_{it}$ is a vector of covariates and γ is the vector of corresponding regression coefficients. The covariates $\mathbf{w}_{it}$ may be time-independent or time-dependent. From (2.5), the survival function of T_i^* is

$$S_i(t) = \exp\left(-\int_0^t h_0(s) \exp\{\gamma' \mathbf{w}_{is} + \alpha\eta_{is}\} ds\right). \tag{2.6}$$

In this model, the association between the claim payment process and the settlement process is introduced through the effects of η_{it} on the hazard of settlement that is measured by α. A positive α indicates a negative payment-settlement relation, i.e. larger payments will accelerate the settlement, and vice versa. For the baseline hazard in (2.5) both the Weibull model and an approximation based on splines are considered. The Weibull baseline is given by:

$$h_0(t) = \lambda k t^{k-1}, \tag{2.7}$$

where λ is the scale parameter, and k is the shape parameter. When $k = 1$, $h_0(t)$ reduces to an exponential baseline function. The Weibull model is commonly used because of its simplicity and easy interpretability. However, with only two parameters, the Weibull model has limited flexibility for fitting different baseline hazard functional forms. A more flexible model is to approximate the baseline hazard using splines. Splines are piecewise polynomials satisfying continuity constraints at the knots joining the pieces (Gray, 1992). Specifically, consider:

$$\log h_0(t) = \lambda_0 + \sum_{k=1}^{K} \lambda_k B_k(t, q). \tag{2.8}$$

Here, $B_k(\cdot)$ is a B-spline basis function, q denotes the degree of the B-spline basis function, $K = q + m$; where m is the number of interior knots, and $\lambda = (\lambda_0, \lambda_1, \cdots, \lambda_K)$ are the spline coefficients. The spline provides great flexibility as the number of knots increases. But, preferences of the number of knots and their locations are important to the overall fit. For convenience, I denote ω to be the unknown parameters in the baseline hazard model.

2.5 Statistical Inference

2.5.1 Estimation

The parameters of the joint model are estimated using likelihood-based methods. Denote $\theta = (\theta_{\mathbf{1}}, \theta_{\mathbf{2}})$, where $\theta_{\mathbf{1}}$ summarizes the parameters of the longitudinal submodel including both regression coefficients β and variance components (ν, ϕ), and $\theta_{\mathbf{2}}$ summarizes the parameters of the survival submodel that includes baseline hazard ω, regression coefficients γ, and association between claim payment and settlement α. The likelihood function for the observables $(t_i, \delta_i, \mathbf{y}_i)$ of claim i is shown as:

$$\begin{aligned} L(\theta; t_i, \delta_i, \mathbf{y}_i) &= \int f(\mathbf{y}_i|\mathbf{b}_i; \theta) f(t_i, \delta_i|\mathbf{b}_i; \theta) dF(\mathbf{b}_i; \theta). \\ &= \int \left[\prod_{t\in\tau_i} f(y_{it}|\mathbf{b}_i; \theta)\right] f(t_i, \delta_i|\mathbf{b}_i; \theta) f(\mathbf{b}_i; \theta) d\mathbf{b}_i, \end{aligned} \tag{2.9}$$

where

$$\begin{aligned} f(t_i, \delta_i|\mathbf{b}_i; \theta) &= (h_i(t_i|\mathbf{b}_i; \theta))^{\delta_i} S_i(t_i|\mathbf{b}_i; \theta) \\ &= (h_0(t_i) \exp\{\gamma' \mathbf{w}_{it_i} + \alpha\eta_{it_i}\})^{\delta_i} \exp\left(-\textstyle\int_0^{t_i} h_0(s) \exp\{\gamma' \mathbf{w}_{is} + \alpha\eta_{is} ds\}\right). \end{aligned} \tag{2.10}$$

Given data collected on N individual claims, the MLE of model parameters are obtained by:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \log L(\theta; t_i, \delta_i, \mathbf{y}_i). \tag{2.11}$$

The variance of $\hat{\theta}$ is estimated using the inverse of the observed Information matrix, i.e. $Var(\hat{\theta}) = [I(\hat{\theta})]^{-1}$, where

$$I(\hat{\theta}) = -\sum_{i=1}^{N} \frac{\partial^2 \log L(\theta; t_i, \delta_i, \mathbf{y}_i)}{\partial\theta\partial\theta'}\Big|_{\theta=\hat{\theta}} \quad , \tag{2.12}$$

and the second order derivative is approximated by the numerical Hessian matrix. I employ a normal random effects in the joint model. Song et al. (2002) proposed an estimation procedure that does not require normality assumption for random effects and showed that estimation under normal normal assumption is robust to misspecification. In addition, Rizopoulos et al. (2008) showed that misspecification of the random effects distribution has a minimal effect in parameter estimation that wanes when the number of repeated measurements increases.

Evaluation of the likelihood function is computationally difficult because of the integral in the likelihood function (2.9) and the integral in the survival density function (2.10). Numerical integration techniques such as Gaussian quadrature (Song et al., 2002), Monte Carlo (Henderson et al., 2000) and Laplace approximations (Rizopoulos et al., 2009) have been applied in the joint modeling framework. Maximization approaches include the EM algorithm that treats the random effects as missing data (Wulfsohn and Tsiatis, 1997) and a direct maximization of the log-likelihood using a quasi-Newton algorithm (Lange, 2004). In this study, the Gaussian quadrature numerical techniques is employed to evaluate the likelihood function.

For claim specific predictions, in addition to the joint model's MLE estimates $\hat{\theta}$, an estimate of the random-effects is needed. The random-effects estimate $\hat{\mathbf{b}}_i$ is obtained using Bayesian methods with posterior distribution:

$$f(\mathbf{b}_i|t_i,\delta_i,\mathbf{y}_i;\hat{\theta}) = \frac{f(t_i,\delta_i|\mathbf{b}_i;\hat{\theta})f(\mathbf{y}_i|\mathbf{b}_i;\hat{\theta})f(\mathbf{b}_i|\hat{\theta})}{f(t_i,\delta_i,\mathbf{y}_i;\hat{\theta})}. \tag{2.13}$$

Then the empirical Bayes estimate is obtained using the mean $\hat{\mathbf{b}}_i$ of the posterior distribution given by;

$$\hat{\mathbf{b}}_i = \int \mathbf{b}_i f(\mathbf{b}_i|t_i,\delta_i,\mathbf{y}_i;\hat{\theta})d\mathbf{b}_i. \tag{2.14}$$

2.5.2 Prediction

For the prediction of unpaid losses, the focus is on open claims at the valuation time. An open claim at valuation time is characterized by time since reporting c_i, longitudinal claim history $\mathcal{Y}_i(c_i) = \{y_{it}, 0 \leq t \leq c_i\}$ and implies that the settlement time $T_i^* > c_i$. With the fitted joint model, the RBNS reserve prediction for the ith claim at the valuation time, $\hat{R}_i^{RBNS}(c_i)$, can be obtained using the following steps which are elaborated in Algorithm 1:

a) Predict the future time when the ith claim will be settled, $\hat{u}_i$, given $T_i^* > c_i$ and $\mathcal{Y}_i(c_i)$ using (2.19) from Section 2.5.2.1
b) Predict the ultimate payment, $\hat{Y}_i^{ULT}(\hat{u}_i)$, given $\mathcal{Y}_i(c_i)$ and $T_i^* > c_i$ using (2.20) from section 2.5.2.2.
c) With the cumulative payment for the ith claim at valuation time, $Y_i(c_i)$, then:

$$\hat{R}_i^{RBNS}(c_i) = \hat{Y}_i^{ULT}(\hat{u}_i) - Y_i(c_i). \tag{2.15}$$

Let m be the number of open claims at the valuation time, i.e. $m = \sum_{i=1}^{N} I(\delta_i = 0)$. Then the total RBNS reserve amount is given by:

$$\hat{R}^{RBNS}(c) = \sum_{i=1}^{m} \hat{R}_i^{RBNS}(c_i). \tag{2.16}$$

2.5.2.1 Prediction of Time-to-Settlement

To predict the time-to-settlement for a RBNS claim, given that the claim survived (not settled) up to the valuation time, we are interested in estimating the conditional survival probability:

$$\pi_i(u|c_i) = \Pr(T_i^* \geq u | T_i^* > c_i, \mathcal{Y}_i(c_i); \theta) = \frac{S_i\left(u|\eta_{iu}, \mathbf{w}_{iu}; \theta\right)}{S_i\left(c_i|\eta_{ic_i}, \mathbf{w}_{ic_i}; \theta\right)}, \tag{2.17}$$

where $S_i(\cdot)$ is given by (2.6), and $u > c_i$. $\mathbf{w}_{iu}$ and $\mathbf{w}_{ic_i}$ are covariates at times u and

c_i. $\pi_i(u|c_i)$ gives the probability that there are further payments at future time u. Here, the probability prediction is dynamic because $\pi_i(u|c_i)$ depends on the expected claims amounts at valuation time c_i and future time u given by η_{ic_i} and η_{iu}, respectively. Then the predictions can be updated as more data becomes available. Using the MLE estimates $\hat{\theta}$ and the empirical Bayes estimate $\hat{\mathbf{b}}_i$, an estimate of $\pi_i(u|c_i)$ is given by:

$$\hat{\pi}_i(u|c_i) = \frac{\hat{S}_i\left(u|\hat{\eta}_{iu}, \mathbf{w}_{iu}; \hat{\theta}\right)}{\hat{S}_i\left(c_i|\hat{\eta}_{ic_i}, \mathbf{w}_{ic_i}; \hat{\theta}\right)}, \tag{2.18}$$

where $\hat{\eta}_{iu} = \mathbf{x}'_{iu}\hat{\beta} + \mathbf{z}'_{iu}\hat{\mathbf{b}}_i$ and $\hat{\eta}_{ic_i} = \mathbf{x}'_{ic_i}\hat{\beta} + \mathbf{z}'_{ic_i}\hat{\mathbf{b}}_i$. The time-to-settlement for a RBNS claim, $\hat{u}_i = E(T_i^*|T_i^* > c_i, \mathcal{Y}_i(c_i))$ is given by:

$$\hat{u}_i = \int_{c_i}^{\infty} \hat{\pi}_i(u|c_i)du. \tag{2.19}$$

2.5.2.2 Prediction of Future Claim Payments

For the future claim payments prediction of an open claim at the valuation time, we are interested in the expected cumulative payments at future time $u > c_i$ for the ith claim conditional on longitudinal claim history $\mathcal{Y}_i(c_i)$, $E\left[Y_i(u)|T_i^* > c_i, \mathcal{Y}_i(c_i)\right]$, given by:

$$\hat{Y}_i(u) = g^{-1}(\mathbf{x}'_{iu}\hat{\beta} + \mathbf{z}'_{iu}\hat{\mathbf{b}}_i). \tag{2.20}$$

Here, $g^{-1}(\cdot)$ is the inverse of the link function, $\{\mathbf{x}_{iu}, \mathbf{z}_{iu}\}$ are covariates, and $\hat{\beta}$ are the maximum likelihood estimates. When $\hat{u}_i$ is the estimated time-to-settlement for ith claim, $\hat{Y}_i(\hat{u}_i)$ is the predicted ultimate amount of the claim.

An algorithm for predicting the loss reserve using the joint model is given in Algorithm 1.

Algorithm 1 Reserve prediction routine for JM.

Input: Valuation time c_i, observed data $(t_i, \delta_i, \mathbf{y}_i, \mathbf{w_{ic_i}}, \mathbf{x_{ic_i}}, \mathbf{z_{ic_i}})$, MLE $\hat{\theta}$, $\hat{Var}(\hat{\theta})$, empirical Bayes estimate $\hat{\mathbf{b}}_i$, cumulative amount paid $Y_i(c_i)$, future time u, covariates at time u $(\mathbf{w_{iu}}, \mathbf{x_{iu}}, \mathbf{z_{iu}})$, and number of draws K.

Output: $\left\{\hat{R}_i^{RBNS}(c_i); i = 1, ..., m\right\}$;

for $i = 1, ..., m$ **do**

  Calculate $\hat{\eta}_{ic_i} = \mathbf{x}'_{ic_i}\hat{\beta} + \mathbf{z}'_{ic_i}\hat{\mathbf{b}}_i$;

  Calculate $\hat{S}_i(c_i) = \exp\left(-\int_0^{c_i} \hat{h}_0(s)\exp\{\hat{\gamma}'\mathbf{w}_{is} + \hat{\alpha}\hat{\eta}_{is}\}ds\right)$;

  for $k = 1, ..., K$ **do**

    Generate $\hat{\pi}_i(u|c_i) = U \sim \text{Uniform}(0, 1)$;

    Calculate $u_{ik} = \hat{H}_i^{-1}(-\log(U \times \hat{S}_i(c_i)))$; $\hat{H}_i(u) = \int_0^u \hat{h}_0(s)\exp\{\hat{\gamma}'\mathbf{w}_{is} + \hat{\alpha}\hat{\eta}_{is}\}ds$;

  end for

  return $\{u_{ik}; k = 1, ..., K\}$;

  Calculate $\hat{u}_i = K^{-1}\sum_{k=1}^K u_{ik}$;

  Generate $\hat{Y}_i^{ULT_l}(\hat{u}_i) = g^{-1}(\hat{\eta}_{i\hat{u}_i})$; $\hat{\eta}_{i\hat{u}_i} = \mathbf{x}'_{i\hat{u}_i}\hat{\beta} + \mathbf{z}'_{i\hat{u}_i}\hat{\mathbf{b}}_i$;
  For parameter uncertainty.

  Generate $\hat{Y}_i^{ULT_l}(\hat{u}_i) \sim f\left(g^{-1}(\hat{\eta}_{i\hat{u}_i}); \hat{\theta}\right)$;
  For parameter and process uncertainty.

  Calculate $\hat{R}_i^{RBNS}(c_i) = \hat{Y}_i^{ULT}(\hat{u}_i) - Y_i(c_i)$;

  return $\left\{\hat{R}_i^{RBNS}(c_i); i = 1, ..., m\right\}$;

end for

Chapter 3

Evaluating the Joint Model Framework Using Simulated Data

Chapter Preview. To better understand the strengths and limitations of the joint model framework, this chapter investigates the estimation and prediction performance of the joint model framework described in Chapter 2 using simulated data.

Section 3.1 describes the design for the simulation study. Section 3.2 emphasizes the importance of the joint model on parameter estimation, and Section 3.3 evaluates the prediction performance of the joint model. Section 3.4 identifies environmental changes where the joint model outperforms macro-level reserving methods. Section 3.5 discusses the parameter and process uncertainty components of the prediction distribution of the joint model. Section 3.6 concludes.

3.1 Simulation Design

In the simulation, the longitudinal submodel is assumed to be a gamma regression with dispersion parameter $1/\sigma$. The conditional mean is further specified as

$$\eta_{it} = g(E[Y_{it}|\mathbf{b}_i]) = \mathbf{x}'_{it}\beta + \mathbf{z}'_{it}\mathbf{b}_i = \beta_{10} + t\beta_{11} + x_{i1}\beta_{12} + x_{i2}\beta_{13} + b_{i0}. \tag{3.1}$$

The survival submodel is a proportional hazards model with an exponential base hazard. Specifically, the conditional hazard function is:

$$h_i(t|\eta_{it}, \mathbf{w}_{it}) = h_0(t)\exp\{\gamma_1 x_{i1} + \gamma_2 x_{i2} + \alpha\eta_{it}\} \quad \text{and} \quad h_0(t) = \lambda, \tag{3.2}$$

where $\mathbf{x}_{it} = \{t, x_{i1}, x_{i2}\}$ and $\mathbf{w}_{it} = \{x_{i1}, x_{i2}\}$. The random effects are generated from a normal distribution $\mathcal{N}(0, \nu)$.

The parameters used in data generation are summarized in Table 3.1. The payment times are assumed to be exogenous and are set at $t = 0, 1, 2, ..., 9$. I assume $x_1 \sim Bernouli(0.5)$, representing a discrete predictor and $x_2 \sim Normal(1, 0.5)$, corresponding to a continuous predictor. The claims are evenly and independently distributed among ten accident years, and the censoring time is the end of calender year ten. Based on the work of Sweeting and Thompson (2011), I employ Algorithm 2 to construct the training and validation data in the simulation study, and a sample R code is given in Appendix 7.3.

3.2 Parameter Estimates

The main results on parameter estimation are summarized in Table 3.1. I consider different sample sizes (number of claims) and report the results for N=500, 1000, and 1500. For each simulated sample, the parameter estimates and the associated standard error are obtained using the likelihood-based method describe in Section 2.5.1. The results reported in Table 3.1 are based on S=150 replications.

I show in the table the average bias (Bias) and the average standard error (SE) of the estimates. In addition, I calculate the nominal standard deviation of the point estimates (SD), and report the standard deviation of the average bias calculated as

Algorithm 2 Data-generating process for JM.

Input: Parameters $\{\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \nu, \sigma\}$ from the payments submodel, and $\{\gamma_1, \gamma_2, \lambda, \alpha\}$ from the settlement submodel.

Output: Training datasets $D_T^P = \{(y_{it}, t, x_{i1}, x_{i2}); 0 \leq t \leq t_i, i = 1, ..., N\}$, and $D_T^S = \{(t_i, \delta_i, x_{i1}, x_{i2}); i = 1, ..., N\}$; Validation dataset for open claims $D_V = \{y_{it}; c_i < t \leq t_i^*, i = 1, ..., m\}$.

for Claim $i = 1, ..., N$ **do**
  Generate $x_{i1} \sim$ Bernouli(0.5), $x_{i2} \sim$ Normal$(1, 0.5)$;
  Generate $\mathbf{b}_i = b_{i0} \sim \mathcal{N}(\mathbf{0}, \nu)$;
  for Payment time $t = 0, ..., 9$ **do**
    $y_{it} \sim$ Gamma $\left(\frac{\exp(\eta_{it})}{\sigma}, \sigma\right)$; $\eta_{it} = \beta_{10} + t\beta_{11} + x_{i1}\beta_{12} + x_{i2}\beta_{13} + b_{i0}$;
  end for
  return $\{y_{it}; t = 0, ..., 9\}$;
  Generate $S_i(t) = U \sim$ Uniform$(0, 1)$;
  Calculate $t_i^* = H_i^{-1}(-\log(U))$; where $H_i(t) = \int_0^t \lambda \exp\{\gamma_1 x_{i1} + \gamma_2 x_{i2} + \alpha\eta_{is}\} ds$;
  Generate accident year $AY_i \in [(1, \ldots, 10) - 1]$;
  Generate $c_i = 9 - AY_i +$ Uniform$(0, 0.5)$;
  $D_{T\,i}^P = \{(y_{it}, t, x_{i1}, x_{i2}); 0 \leq t \leq t_i\}$; where $t_i = \min(t_i^*, c_i)$;
  $D_{T\,i}^S = \{t_i, \delta_i, x_{i1}, x_{i2}\}$; where $\delta_i = I(t_i^* < c_i)$;
  $D_{V\,i} = \{y_{it}; c_i < t \leq t_i^*, \delta_i = 0\}$;
  return $D_T^P = \left\{D_{T\,i}^P; i = 1, ..., N\right\}$; $D_T^S = \left\{D_{T\,i}^S; i = 1, ..., N\right\}$; and $D_V = \{D_{V\,i}; i = 1, ..., m\}$; where $m = \sum I(\delta_i = 0)$;
end for

SD/$\sqrt{S}$. As anticipated, both estimate and uncertainty of the average bias decrease as sample size increases. The average standard error is comparable to the nominal standard deviation, indicating the accuracy of variance estimates. Lastly, the standard errors are consistent with $\sqrt{n}$ convergence. The estimation results in Table 3.1 show that the bias in estimating parameters from the joint model is negligible.

Table 3.1: Estimation results for JM for different sample sizes (number of claims).

S=150		Bias			SD/$\sqrt{S}$			SE	
Parameter	N=500	1000	1500	500	1000	1500	500	1000	1500
Longitudinal submodel(GLMM)									
β_{10}=1.0	0.003	0.001	-0.008	0.005	0.005	0.004	0.059	0.056	0.051
β_{11}=0.3	0.001	0.002	0.001	0.001	0.001	0.001	0.011	0.010	0.010
β_{12}=0.2	-0.008	-0.001	0.001	0.004	0.003	0.004	0.053	0.039	0.042
β_{13}=0.4	-0.002	-0.002	0.006	0.004	0.003	0.003	0.044	0.042	0.039
ν=0.09	0.000	-0.001	0.000	0.001	0.001	0.001	0.018	0.015	0.016
σ=1.5	0.004	0.001	0.005	0.005	0.004	0.003	0.055	0.043	0.038
Survival submodel									
γ_1=0.5	0.000	-0.004	0.000	0.008	0.007	0.007	0.101	0.085	0.081
γ_2=0.3	0.007	-0.001	0.000	0.009	0.007	0.006	0.106	0.079	0.078
log(λ)=-1.139	-0.036	-0.021	-0.012	0.015	0.012	0.013	0.181	0.148	0.153
α=-0.25	0.010	0.011	0.005	0.007	0.005	0.006	0.083	0.066	0.078

To emphasize the importance of joint estimation, I explore two additional estimation strategies, independent and two-stage estimations. The former estimates the longitudinal and survival submodels separately ignoring the association between the two components, i.e., setting $\alpha = 0$ in the survival submodel. The latter estimates the parameters in the longitudinal submodel in the first stage, and then estimates the parameters in the survival submodel in the second stage holding the longitudinal model parameters fixed. See Appendix 7.2 for details on these estimation techniques. It turns out that both estimation strategies introduce significant bias in the parameter estimates. Estimation results based on sample size N=1000 and S=150 replications are reported in Table 3.2 and Table 3.3, respectively, for the longitudinal and survival submodels. I show in the table the average bias (Bias), the average standard error (SE) of the estimates, and the standard deviation of the average bias calculated as SD/$\sqrt{S}$. For comparison, I reproduce the estimates for the joint model from Table 3.1.

It is critical to note that both estimation strategies induce substantial bias into parameter

estimates. For the independence method, I emphasize that it is different from the usual multivariate regression where ignoring the association among multiple outcomes pays no price in terms of consistency, but only hampers the efficiency. The bias in the longitudinal submodel is due to the sample selection under independence assumption and the bias in the survival submodel is due to the omitted variable. For the two-stage estimation, the selection bias in the longitudinal submodel is carried over to the survival submodel. Therefore, model parameters cannot be consistently estimated although the association between the two processes is taken into account.

Table 3.2: Estimation results for the longitudinal submodel.

N=1000, S=150	JM			Independence and Two-stage		
Parameter	Bias	SD/$\sqrt{S}$	SE	Bias	SD/$\sqrt{S}$	SE
β_{10}=1.0	0.001	0.005	0.056	-0.035	0.005	0.056
β_{11}=0.3	0.002	0.001	0.010	0.008	0.001	0.010
β_{12}=0.2	-0.001	0.003	0.039	0.001	0.003	0.040
β_{13}=0.4	-0.002	0.003	0.042	-0.001	0.003	0.042
ν=0.09	-0.001	0.001	0.015	0.002	0.001	0.017
σ=1.5	0.001	0.004	0.043	0.092	0.006	0.072

Table 3.3: Estimation results for the survival submodel.

N=1000, S=150	JM			Two-Stage			Independence		
Parameter	Bias	SD/$\sqrt{S}$	SE	Bias	SD/$\sqrt{S}$	SE	Bias	SD/$\sqrt{S}$	SE
γ_1=0.5	-0.004	0.007	0.085	-0.004	0.007	0.086	-0.026	0.007	0.089
γ_2=0.3	-0.001	0.007	0.079	-0.004	0.007	0.080	-0.084	0.007	0.081
log(λ)=-1.139	-0.021	0.012	0.148	-0.075	0.012	0.140	-0.420	0.009	0.116
α=-0.25	0.011	0.005	0.066	0.020	0.005	0.063	-	-	-

3.3 RBNS Prediction

This section focuses on the prediction performance of the proposed joint model in different scenarios. The prediction from the joint model is compared with results from the independent and two-stage estimation techniques. The results presented in this section are based on a sample size of $N = 1000$ and $S = 150$ replications. The predictive routine used for the joint model is elaborated in Algorithm 1; the predictive routine for the independent and two-stage techniques is similar to that of the joint model.

3.3.1 High Frequency Versus Low Frequency Payments

In this scenario, I investigate the effect of payment frequency from individual claims on the prediction accuracy. The payment frequency is defined as the number of payments per unit time period. The high-frequency payment case corresponds to the base model described in Section 3.1 where the maximum number of payments for each claim is ten, and payments are at times $t = 0, 1, 2, ..., 9$. In the low-frequency payment case, the maximum number of payments is reduced by half, and payment times are $t = 0, 2, 4, 6, 8$. Note that the payment frequency does not alter the settlement process, and it only affects the number of observations generated from the longitudinal submodel.

Figure 3.1 illustrates the timeline of the payment times for the low-frequency and high-frequency payment scenarios. It is seen that claims in the high-frequency scenario are likely to have more payment transactions than those in the low-frequency scenario. For instance, a claim that is to be settled at $t = 1.5$ will be closed with a single transaction under the low-frequency payment scenario. However, a claim with the same settlement time will be closed with two transactions under the high-frequency payment scenario.

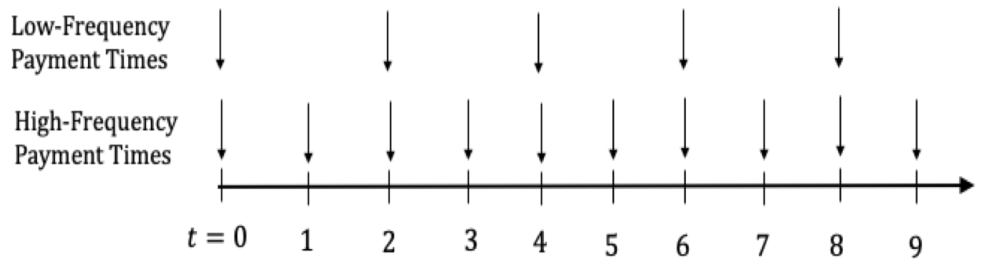

Figure 3.1: Payment times for low-frequency and high-frequency payment scenarios.

One can think of the low-frequency payment scenario as a representation of short-tail business lines such as personal automobile collision insurance, where claims, once reported, are typically settled with a single payment within a relatively short period of time. In contrast, the high-frequency payment scenario mimics long-tail business lines such as workers' compensation insurance, where claim settlement is usually accompanied

by more payment transactions than the short-tail lines.

Table 3.4 shows the true RBNS reserve, the reserve error (estimated RBNS reserve minus the actual unpaid losses), the error as a percentage of actual unpaid losses, and the standard error of prediction divided by the number of replications (SE/$\sqrt{S}$). For the high-frequency scenario simulation, it is seen that JM performs better than the independent and the two-stage estimation techniques with the smallest percentage error of 0.55%. The performance of the two-stage technique and independent technique in comparison to the JM model emphasizes the point that when the endogenous nature of the cumulative payments and the association between cumulative payments and settlement process are ignored, it leads to biased estimates and consequently inaccurate predictions of unpaid losses.

For the low-frequency simulation, JM with the percentage error of 1.16% again performs better than the other estimation techniques. The slight increase in the percentage reserve error for the two-stage and the joint model compared to the high-frequency model indicates that the reduction in payment transactions reduces the accuracy of the individual claim random effects estimate used for the reserve predictions. Also, compared to the high-frequency model, the percentage reserve error for the independent model has reduced to -20.58%, which implies that the advantages of the joint model are significant for long-tail lines of business.

Table 3.4: RBNS prediction results under high and low frequency payments.

N=1000, S=150	High-Frequency			Low-Frequency		
	Mean	Error %	SE/$\sqrt{S}$	Mean	Error %	SE/$\sqrt{S}$
True Reserve	6,062		71	4,412		49
JM Error	33	0.55	74	51	1.16	55
Two-Stage Error	206	3.39	77	208	4.72	59
Independent Error	-1,583	-26.12	57	-908	-20.58	45

3.3.2 Model Misspecifications

Correct model specification is crucial to accurate reserving prediction. This section examines the prediction performance of the joint model when either the longitudinal submodel or the survival submodel is misspecified. In particular, I study the impact of the misspecification of the payment trend in the longitudinal submodel and the misspecification of the baseline hazard in the survival submodel.

To investigate the effects of misspecification of the longitudinal submodel on the prediction performance of the JM, a longitudinal cumulative payments with quadratic payment trend is simulated. The simulated data is fitted assuming a linear trend. That is, the true cumulative payment model is $\eta_{it} = \beta_{10} + t\beta_{11} + t^2\beta_{1*} + x_{i1}\beta_{12} + x_{i2}\beta_{13} + b_{i0}$ and the misspecified fitted payment model is $\eta_{it} = \beta_{10} + t\beta_{11} + x_{i1}\beta_{12} + x_{i2}\beta_{13} + b_{i0}$. The left panel of Figure 3.2 shows the true cumulative trend and fitted trend. The true payment model has the same parameters from the base model in Table 3.1, and also contains an extra term for the square of the payment time variable (t^2) with coefficient $\beta_{1*} = -0.003$. The survival submodel also has the same parameters from the base model in Table 3.1.

I also investigate the impact of misspecification of the baseline hazard in the survival submodel on the prediction performance of the joint model. Here, the true baseline hazard function follows a Weibull distribution from (2.7) where $k = 1.1$ but the fitted model assumes an exponential baseline with $k = 1$. The longitudinal submodel and regression coefficients of the survival submodel has the same parameters from the base model in Table 3.1. The right panel of Figure 3.2 shows the true and misspecified baseline hazard.

Table 3.5 shows the prediction results when there is misspecification of the payment trend in the longitudinal submodel and also misspecification of the baseline hazard in the survival submodel. Using a linear trend in the misspecified longitudinal submodel as seen in Figure 3.2 will lead to more significant ultimate claim amount prediction, hence overstate the reserve prediction, and that is confirmed in Table 3.5.

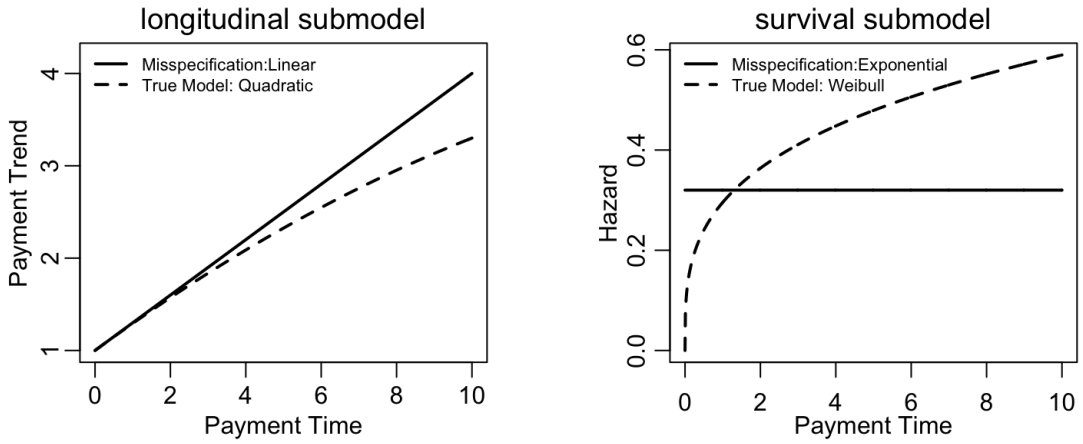

Figure 3.2: Misspecification of longitudinal and survival submodels.

Table 3.5: RBNS prediction results with model misspecifications.

N=1000, S=150	Misspecified longitudinal submodel			Misspecified survival submodel		
	Mean	Error %	SE/$\sqrt{S}$	Mean	Error %	SE/$\sqrt{S}$
True Reserve	4,623		54	4,306		52
JM Error	576	12.45	62	-618	-14.35	60
Two-Stage Error	715	15.46	66	-599	-13.92	64
Independent Error	-769	-16.63	48	-1070	-24.85	52

For this specific example, the misspecification of the longitudinal submodel's payment trend produced a prediction error of 12.45%. This performance underscores the importance of the correct specification of the underlying longitudinal submodel. The results imply that for application on real-world data, the joint model diagnostics should focus not only on the underlying distribution of the longitudinal submodel but also on the payment trend. Dean (2014) discusses several statistics for comparing nested and non-nested models.

In the same way, the poor prediction performance under the misspecification of the baseline hazard in the survival submodel emphasizes the importance of the correct specification of the underlying submodels. For real-world data applications, to evaluate the goodness of fit of the survival function, the Cox-Snell residuals r_i can be employed. The Cox-Snell residual is calculated as the value of cumulative hazard function $H(T_i)$ evaluated at the observed event time T_i and given by

$$r_i = \int_0^{T_i} h_i(s)ds. \tag{3.3}$$

Here, the Kaplan-Meier estimate of the survival function of r_i is compared to the survival function of unit exponential distribution (Rizopoulos, 2012), i.e., if the model fits the data well, r_i is expected to have a unit exponential distribution. Given, $T_i \sim S(t_i)$ then $S(T_i) \sim$ Uniform(0,1), and $r_i = H(T_i) = -\log(S(T_i)) \sim$ Exponential (1). The survival model in Figure 3.3 was simulated using a Weibull baseline hazard model. The solid line is the Kaplan-Meier estimate of the survival function of r_i, and the dashed line is the survival function of the unit exponential distribution. It can be seen that both the Weibull and spline baseline fits the data very well. This shows that the spline baseline model is flexible.

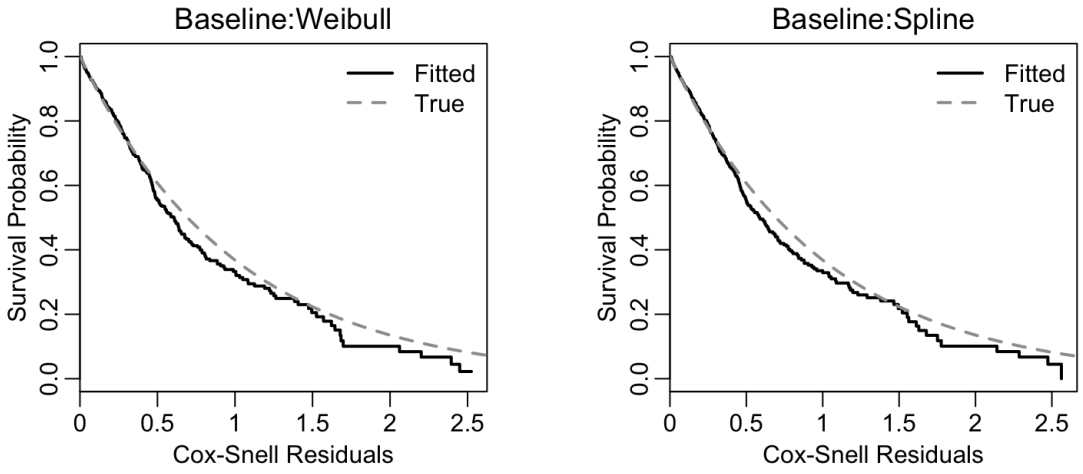

Figure 3.3: Evaluation of the survival model fit using Cox-Snell residuals.

3.4 Environmental Changes

This section investigates the effects of environmental changes on reserving prediction. It is well known that the industry benchmark chain-ladder method relies on a stable environment, and is expected to fail when the insurer undergoes significant operational changes that change the claim development pattern. In contrast, individual-level reserving methods leverage granular claims level data and are thus capable of capturing such changes and reflecting them in predicting unpaid losses. I show that the proposed joint model can easily accommodate environmental changes that affect reserving prediction. I consider various examples of environmental changes including changes in the underwriting criteria, claim processing, and product mix.

I also provide results using trended chain-ladder techniques based on the recommendation from Berquist and Sherman (1977). The trending techniques treat environmental changes as a trend to adjust the development projections. Berquist and Sherman (1977) presented case studies using a portfolio of U.S. medical malpractice insurance and selected trends to adjust for changes in operations based on a review of the insurer and industry's historical experience. Here, I use a simple trending algorithm, following the work of Jin

(2014), to estimate the trend rate and use the estimated trend rate to make appropriate adjustments prior to the application of traditional development techniques. Because the trending techniques are ad hoc and highly dependent on actuaries' judgments, I provide results using two trending techniques. The first approach assumes that the actuaries are sure of the type of trend that they are dealing with; hence the trending technique accurately captures the environmental changes. Further, the second approach assumes actuaries are not sure of the trend and, therefore, do not accurately capture the environmental changes. See Appendix 7.3.2 for details on the trending techniques.

In the simulation studies, the steady-state is generated from the base model described in Section 3.1. Environmental changes are implemented by using appropriate covariates in either the longitudinal submodel or the survival submodel or both. Table 3.6 provides a brief description of the scenarios that I consider in the numerical experiments. All the prediction results in this section are based on sample size $N = 1000$ and S=150 replications.

Table 3.6: Description of environmental changes and covariates used to implement changes.

Environmental Change	**Description**	**Covariate**
Underwriting Practices	Insurer either tightens or loosens its underwriting criteria due to either changes in the insurer's underwriting guidelines or changes in regulations.	Accident year effects in longitudinal submodel.
Claims Processing	Insurer either speeds up or slows down the claim settlement due to either exterior or interior reasons.	Calendar year effects in survival submodel.
Product Mix	Insurer changes its business mix by increasing or decreasing exposure in long-tail lines of business.	Accident year effects in both longitudinal and survival submodels.

Table 3.7 shows the average of the mean (Mean) and standard deviation (SD) of both settlement time and ultimate paid losses by accident year. As expected in the steady-state, the distribution of both outcomes are stable over time. The results of reserving prediction from both the chain-ladder method and the joint model are displayed in Table 3.8. Not surprisingly, the chain-ladder method performs well under the steady-state with the percentage error of 4.30%, although the joint model produced a superior percentage error of 0.55%. Despite the point prediction from the chain-ladder

and joint model are close, it is worth stressing the difference in the predictive uncertainty from the two models. To illustrate this, I present in the top left panel of Figure 3.4 the predictive distribution of reserve errors. The data aggregation in the chain-ladder leads to information loss, which explains the higher predictive uncertainty compared to the joint model.

Table 3.7: Average of the mean (Mean) and standard deviation (SD) of settlement times and ultimate amount paid under steady-state.

N=1000, S=150		Accident Year									
		1	2	3	4	5	6	7	8	9	10
Settlement Time	Mean	3.08	3.10	3.12	3.12	3.13	3.10	3.07	3.13	3.11	3.11
	SD/$\sqrt{S}$	0.25	0.26	0.26	0.26	0.26	0.26	0.25	0.26	0.26	0.26
Ultimate Payment	Mean	16.32	15.90	15.99	15.79	16.53	16.03	16.07	16.36	16.08	16.16
	SD/$\sqrt{S}$	2.47	2.32	2.35	2.29	2.47	2.37	2.36	2.39	2.37	2.41

Table 3.8: RBNS prediction results under steady-state.

N=1000, S=150	Mean	Error %	SE/$\sqrt{S}$
True Reserve	6,062		71
JM Error	33	0.55	74
Chain-Ladder Error	261	4.30	192

3.4.1 Change in Underwriting Practices

The first scenario of environmental changes that I consider is due to changes in underwriting practices. Insurers use underwriting to evaluate exposures of potential risks, and decide whether the risk is acceptable and how much coverage to provide. The underwriting practice could change due to either changes in the insurer's underwriting guidelines or changes in the regulation. The change in underwriting practice only affects new risks but not existing risks, leading to a shift in the risk profile of the insurer's book. In the reserving context, one would expect a change in loss ratios across accident years.

To implement the change in underwriting practice in simulation, I modify the mean structure of the longitudinal submodel by adding an additional covariate x_3, so that

$$\eta_{it} = g(E[Y_{it}|\mathbf{b}_i]) = \mathbf{x}'_{it}\beta + \mathbf{z}'_{it}\mathbf{b}_i = \beta_{10} + t\beta_{11} + x_{i1}\beta_{12} + x_{i2}\beta_{13} + x_{i3}\beta_{14} + b_{i0}. \quad (3.4)$$

The covariate x_3 is a binary variable that captures the change in the loss ratio across accident years. In the experiment, I set $x_3 = 0$ for accident years 1-5, and $x_3 = 1$ for accident years 6-10, i.e. the shift in loss ratio occurs in accident year 6. The regression coefficient β_{14} controls the direction of the change. A positive (negative) value corresponds to a loosened (tightened) underwriting criteria and thus an increase (decrease) in the loss ratio.

I report the prediction results for $\beta_{14} = 1$. Table 3.9 shows the descriptive statistics of ultimate paid losses and settlement time by accident year. A structural change in the loss amount and the corresponding change in the settlement process over time are observed. Table 3.10 compares the reserve prediction from the chain-ladder method and the joint model. The chain-ladder method does not capture the deteriorating loss ratio in most recent accident years, and thus the projection based on the lower loss ratio significantly underpredicts unpaid losses. In this specific simulation setting, the chain-ladder prediction error is -21.10% . In contrast, actuaries can easily incorporate the information of the change in underwriting in the specification of the joint model and thus adjust for such environmental change in the reserving prediction. Also, though the trended chain-ladder technique (Approach 1), which assumes the actuary is sure of the trend they are dealing with, improved the point estimate, it increased the prediction uncertainty. As expected, the trended chain-ladder technique (Approach 2), which assumes the actuary is not sure of the trend they are dealing with, did not improve the point estimate. The predictive distribution of reserve errors is presented in Figure 3.4, where one observes the bias and high uncertainty in the basic chain-ladder prediction and a higher uncertainty in the trended chain-ladder (Approach 1) prediction.

Table 3.9: Average of the mean (Mean) and standard deviation (SD) of settlement times and ultimate amount paid under change in underwriting practices.

N=1000, S=150		Accident Year									
		1	2	3	4	5	6	7	8	9	10
Settlement Time	Mean	3.08	3.10	3.12	3.12	3.13	3.86	3.83	3.89	3.87	3.87
	SD/$\sqrt{S}$	0.25	0.26	0.26	0.26	0.26	0.29	0.29	0.29	0.29	0.29
Ultimate Payment	Mean	16.32	15.90	15.99	15.79	16.53	55.65	55.54	57.19	54.98	55.50
	SD/$\sqrt{S}$	2.47	2.32	2.35	2.29	2.47	7.95	7.89	8.08	7.80	7.81

I also investigate the case $\beta_{14} = -0.5$ as a result of tightening underwriting criteria. As

Table 3.10: RBNS prediction results under change in underwriting practices.

N=1000, S=150	Mean	Error %	SE/$\sqrt{S}$
True Reserve	19,361		190
JM Error	96	0.50	209
Chain-Ladder Error	-4,085	-21.10	404
Trended CL (Approach 1) Error	324	1.68	748
Trended CL (Approach 2) Error	-3,905	-20.17	419

reported in Table 3.11, the chain-ladder method overestimates the unpaid losses, as anticipated.

Table 3.11: RBNS prediction results under change in underwriting practices (tightening underwriting criteria).

N=1000, S=150	Mean	Error %	SE/$\sqrt{S}$
True Reserve	3,712		52
JM Error	204	5.50	51
Chain-Ladder Error	652	17.56	146
Trended CL (Approach 1) Error	316	8.52	161
Trended CL (Approach 2) Error	610	16.44	147

3.4.2 Changes in Claims Processing

Another common scenario of environmental change relates to the claim service. How claims are handled could be quite different from one insurer to another. Any change in the operation of claim management that affects the speed claims are settled will have a impact on reserving prediction. Such operational changes could be due to both internal or external reasons. For example, a catastrophic loss event could cause a backlog of claims due to short of staffing and thus lead to slowdown in the claim settlement, or an adoption of new information system or technology to streamline claim management that speeds up claim settlement. Another important reason is simply the philosophy in claim processing, for instance, claims could be prioritized based on either their sizes or the order they arrive, and claim adjusters could be assigned based on either the workload or the experience of adjusters.

To reflect the change in claim processing and thus the claim settlement speed, I modify the survival submodel in the data generating process by adding a covariate to indicate

the change. It is worth noting the subtle difference in the effects of change in claim processing and change in underwriting criteria. The difference is in the timing. In a run-off triangle, the change in claim processing affects claims along calender years while the change in underwriting criteria affects claims along accident years; this is because the former applies to both existing and new policies and the latter only applies to new policies. I consider the survival submodel:

$$h_i(t|\eta_{it}, \mathbf{w}_{it}) = h_0(t) \exp\{\gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3t} + \alpha\eta_{it}\}, \tag{3.5}$$

where $x_{i3t} = 0$ if the payment time t is in calendar years 1-5, and $x_{i3t} = 1$ if the payment time t is in calendar years 6-10. The coefficient γ_3 measures the effects on the settlement speed, with a negative value indicating slowdown and a positive value speedup.

To illustrate the change of claim processing, I report in Table 3.12 the descriptive statistics of ultimate paid losses and settlement delay by accident year when setting $\gamma_3 = 0.5$. One observes that the change affects the settlement time, with a negligible impact on the ultimate payments for claims. Table 3.13 reports the corresponding prediction error for RBNS reserves for both chain-ladder and the joint model. Because the chain-ladder assumes the same settlement speed even when the claims are actually closed faster, it overestimates the unpaid losses provided payment pattern stays the same. Again, the trended chain-ladder technique (Approach 1) improved the point estimate and increased the prediction uncertainty. The bottom left panel of Figure 3.4 provides the predictive distributions of reserve errors for the two models and shows a lower predictive uncertainty for the joint model.

Table 3.12: Average of the mean (Mean) and standard deviation (SD) of settlement times and ultimate amount paid under change in claims processing.

N=1000, S=150		Accident Year									
		1	2	3	4	5	6	7	8	9	10
Settlement Time	Mean	2.90	2.81	2.70	2.52	2.28	1.88	1.86	1.90	1.89	1.88
	SD/$\sqrt{S}$	0.23	0.22	0.21	0.19	0.18	0.18	0.17	0.18	0.18	0.17
Ultimate Payment	Mean	14.39	13.64	12.77	11.78	10.72	9.72	9.33	9.58	9.51	9.62
	SD/$\sqrt{S}$	2.03	1.90	1.72	1.62	1.50	1.37	1.28	1.29	1.27	1.32

Table 3.13: RBNS prediction results under change in claims processing.

	Mean	Error %	SE/$\sqrt{S}$
True Reserve	2,256		36
JM Error	-4	-0.19	35
Chain-Ladder Error	1,351	59.90	125
Trended CL (Approach 1) Error	27	1.22	129
Trended CL (Approach 2) Error	1,074	47.62	124

3.4.3 Changes in Product Mix

In the last scenario, I consider the effects of a change in the product mix in the insurer's book on the reserving prediction. Insurance products vary by the nature of the covered risks that could affect both the outstanding payments and the settlement delay. The product mix of an insurer's portfolio could change due to the change in the target markets. For instance, an insurer who provides workers' compensation could shift from low-risk to high-risk occupation class; or a property insurer could decide to expand to write liability insurance; or an insurer could switch target customers from one geographical region to another, etc.

In the simulation, I focus on a situation where the insurer increases exposure in long-tail businesses and reduces exposure in short-tail businesses. Because long-tail lines of business are usually associated with longer settlement and higher losses, I postulate that the change in exposure will increase both outstanding payments and time-to-settlement. To implement this change in the simulation, I use an indicator x_3 to indicate the timing of the change in both the longitudinal and survival submodels as below:

$$\eta_{it} = g(E[Y_{it}|\mathbf{b}_i]) = \mathbf{x}'_{it}\beta + \mathbf{z}'_{it}\mathbf{b}_i = \beta_{10} + t\beta_{11} + x_{i1}\beta_{12} + x_{i2}\beta_{13} + x_{i3}\beta_{14} + b_{i0}. \quad (3.6)$$

$$h_i(t|\eta_{it}, \mathbf{w}_{it}) = h_0(t)\exp\{\gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \alpha\eta_{it}\}. \quad (3.7)$$

The regression coefficients β_{14} and γ_3 quantify the effects on the losses and settlement delay, respectively. Assuming the exposure change takes place in the sixth year, I set $x_3 = 0$ for accident years 1-5, and $x_3 = 1$ for accident years 6-10. Further I set $\beta_{14} = 1$ and $\gamma_3 = -0.5$ to reflect the expectation of larger ultimate losses and longer settlement

time due to increasing exposure in long-tail lines of business.

Table 3.14 summarizes the average of the mean and standard deviation for both ultimate losses and settlement time by accident year. Because of the change in product mix, I observe an increase in both ultimate payments and settlement time starting from the sixth accident year. Note that for simplicity I implement the change as an exogenous shock, i.e. the insurer's new portfolio is formed in the 6th year and is fixed afterwards. A gradual transition to the steady-state of the new portfolio could be easily handled using an interaction with the time trend.

Reserve predictions from the chain-ladder method and the joint model are reported in Table 3.15. Once again, reserving error of the chain-ladder prediction is substantial. Specifically, the chain-ladder underestimates the unpaid loss considerably, because it applies the loss development pattern of short-tailed lines to the business with long-tails without adjustment for the change in the product mix. The trended chain-ladder technique (Approach 1) accurately captures the development pattern changes, hence producing an improved reserve prediction. Because of the limitations of the trending technique, the reserve uncertainty increased significantly. The proposed joint model offers a framework to explicitly accommodate such changes in the model building stage, and thus makes the correction for the product mix change in the reserving prediction as illustrated in the predictive distribution of reserve errors in the bottom right panel of Figure 3.4.

Table 3.14: Average of the mean (Mean) and standard deviation (SD) of settlement times and ultimate amount paid under change in product mix.

N=1000, S=150		Accident Year									
		1	2	3	4	5	6	7	8	9	10
Settlement Time	Mean	3.08	3.10	3.12	3.12	3.13	5.50	5.47	5.53	5.52	5.51
	SD/$\sqrt{S}$	0.25	0.26	0.26	0.26	0.26	0.33	0.33	0.34	0.33	0.33
Ultimate Payment	Mean	16.32	15.90	15.99	15.79	16.53	86.50	85.43	86.57	85.09	84.603
	SD/$\sqrt{S}$	2.47	2.32	2.35	2.29	2.47	10.85	10.66	10.61	10.53	10.24

Table 3.15: RBNS prediction results under change in product mix.

N=1000, S=150	Mean	Error %	SE/$\sqrt{S}$
True Reserve	33,365		234
JM Error	454	1.36	400
Chain-Ladder Error	-16,551	-49.60	446
Trended CL (Approach 1) Error	-1,626	-4.87	1,171
Trended CL (Approach 2) Error	-15,877	-47.59	419

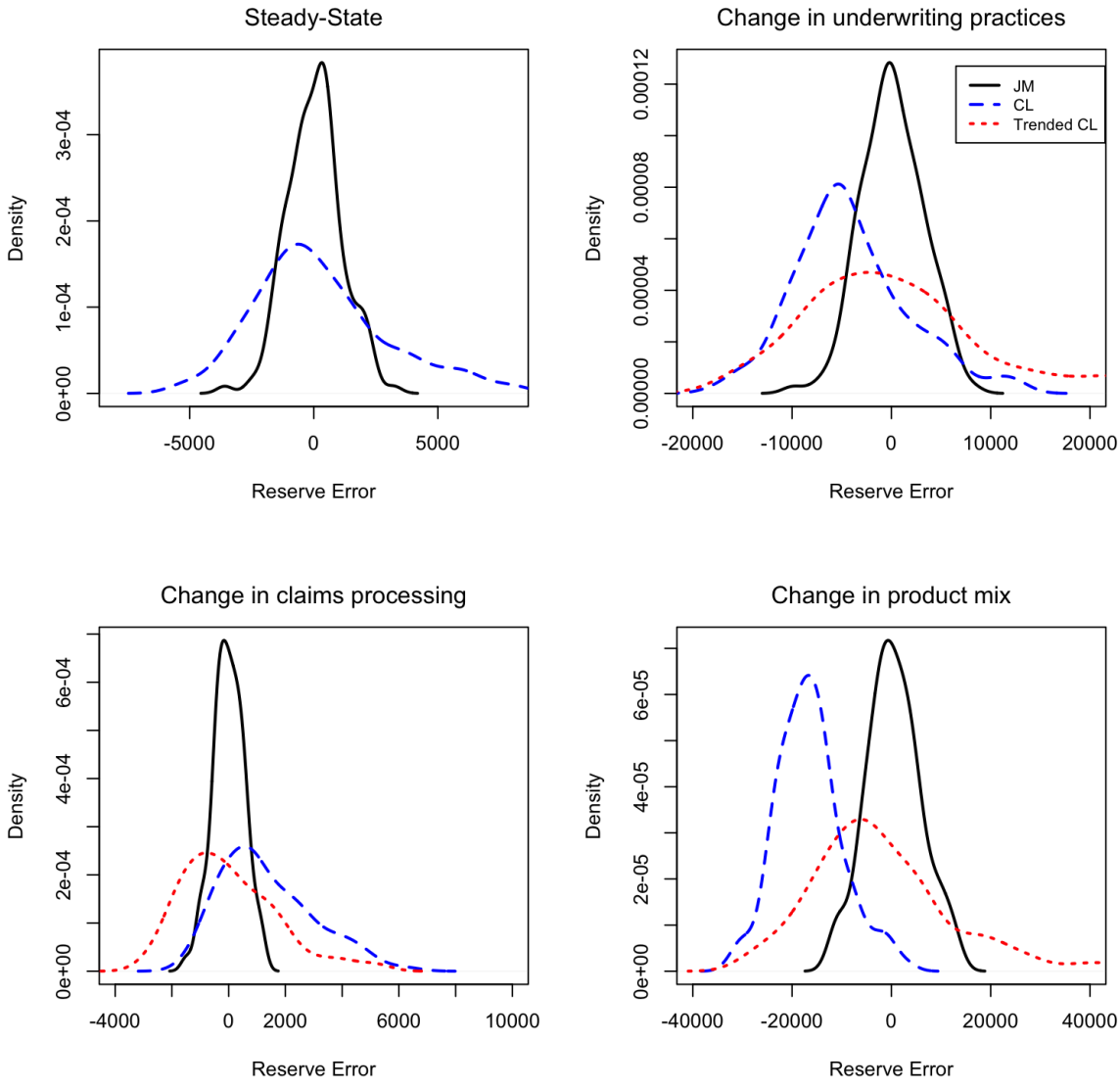

Figure 3.4: Reserve distribution under a steady-state and changes in environmental conditions.

3.5 Parameter and Process Uncertainty

In the previous section, the reserve prediction uncertainty for the joint model was compared to that of the CL method under various scenarios of environmental changes, and the joint model produces lower prediction uncertainty. But so far, the reserve prediction distributions only allowed for parameter uncertainty. As discussed in the first chapter, the mean squared error prediction (MSEP), which provides a measure of the prediction uncertainty, can be decomposed into the parameter uncertainty and the process uncertainty. The parameter uncertainty comes from the uncertainty in the estimation of parameters, and the process uncertainty is the result of the randomness of the development of the claim (England and Verrall, 2002). This section provides the predictive distribution considering both components of the uncertainty.

Figure 3.5 provides the predictive distribution of reserve errors after incorporating both parameter and process uncertainties. For the joint model, the process uncertainty is introduced by simulating the ultimate payment for open claims from a Gamma distribution, as shown in Algorithm 1. For the CL method, to introduce process uncertainty, I use the ODP model and simulate future development from the estimated ODP distribution. There are two important insights from Figure 3.5. First, the CL method still produced biased reserve predictions under unstable environmental conditions as confirmed in tables 3.16 to 3.19. Second, though the joint model produces accurate mean reserve predictions, the reserve uncertainty is higher after incorporating the process uncertainty. The process uncertainty is higher for the joint model because the longitudinal submodel is based on cumulative payments.

Table 3.16: RBNS prediction results under steady-state, after incorporating both parameter and process uncertainty.

N=1000, S=150	Mean	Error %	$\frac{RMSEP}{\sqrt{S}}$
True Reserve	6,062		
JM Error	65	1.07	636
Chain-Ladder Error	141	2.33	201

Further, Figures 3.6 and 3.7 compare the parameter and process uncertainty for the

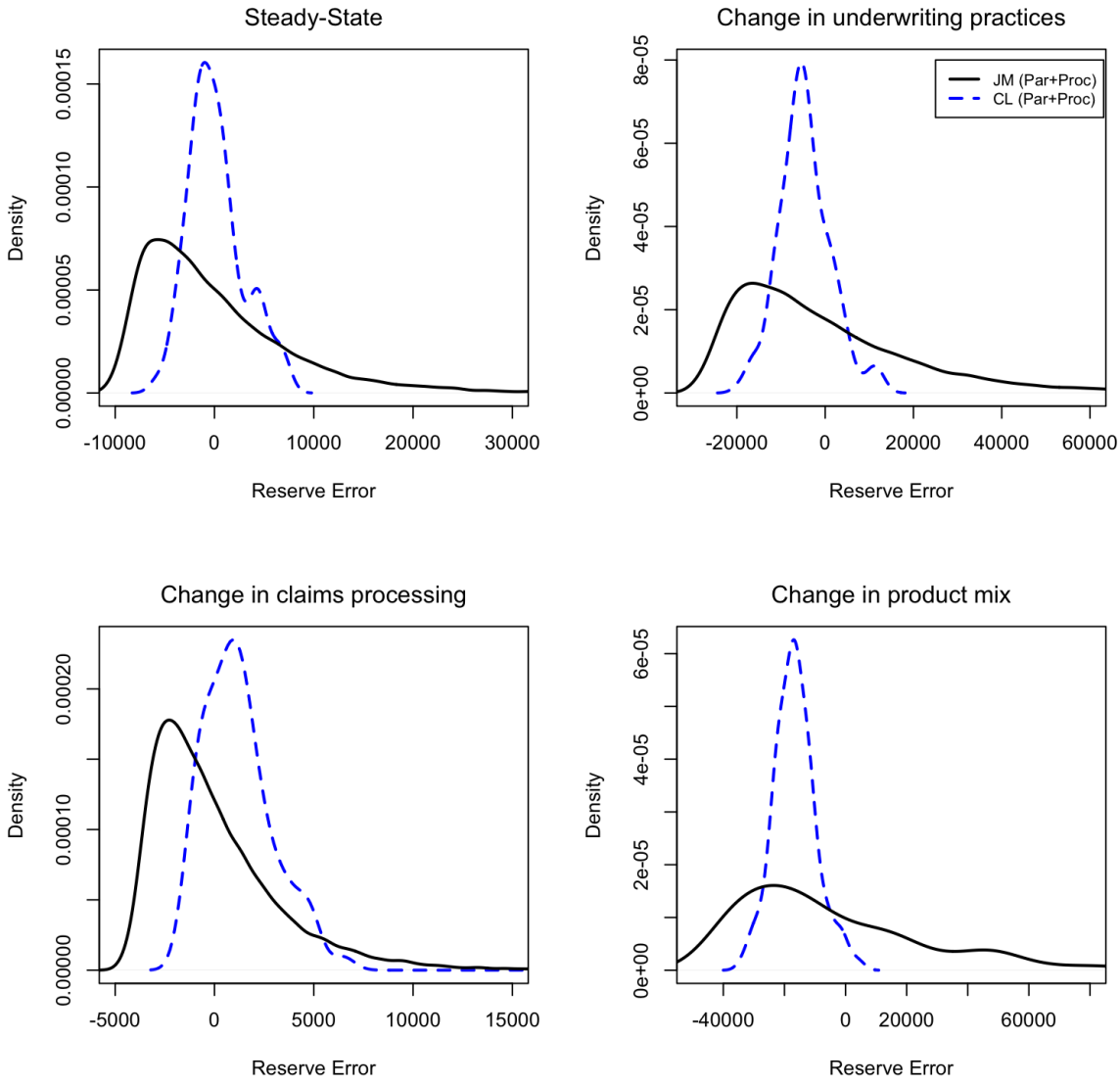

Figure 3.5: Reserve distribution incorporating both parameter and process uncertainty under a steady-state and changes in environmental conditions.

Table 3.17: RBNS prediction results under change in underwriting practices, after incorporating both parameter and process uncertainty.

N=1000, S=150	Mean	Error %	$\frac{RMSEP}{\sqrt{S}}$
True Reserve	19,361		
JM Error	126	0.65	1,800
Chain-Ladder Error	-4,591	-23.71	420

Table 3.18: RBNS prediction results under change in claims processing, after incorporating both parameter and process uncertainty.

	Mean	Error %	$\frac{RMSEP}{\sqrt{S}}$
True Reserve	2,256		36
JM Error	8	0.04	268
Chain-Ladder Error	1,577	69.90	395

Table 3.19: RBNS prediction results under change in product mix, after incorporating both parameter and process uncertainty.

N=1000, S=150	Mean	Error %	$\frac{RMSEP}{\sqrt{S}}$
True Reserve	33,365		234
JM Error	628	1.88	2,952
Chain-Ladder Error	-16,611	-49.79	476

joint model and CL method, respectively, and show that the process uncertainty is a small proportion of reserve uncertainty of the CL method but a significant component of the joint model. The parameter uncertainty is a significant component of the CL predictive distribution because of the limited sample size from the aggregation of data.

3.6 Conclusion

In this chapter, I have demonstrated that failing to incorporate the correlation between the payment processes and the settlement processes could lead to significant error in reserving prediction. Specifically, ignoring the positive (negative) correlation will underpredict (overpredict) the unpaid losses.

In addition, I showed that the proposed joint model could easily accommodate environmental changes such as change in underwriting criteria, business mix, and claim processing among others. Using carefully designed simulation studies, I showed that the industry benchmark chain-ladder method without adjusting for the environmental changes produced substantial error in reserving prediction.

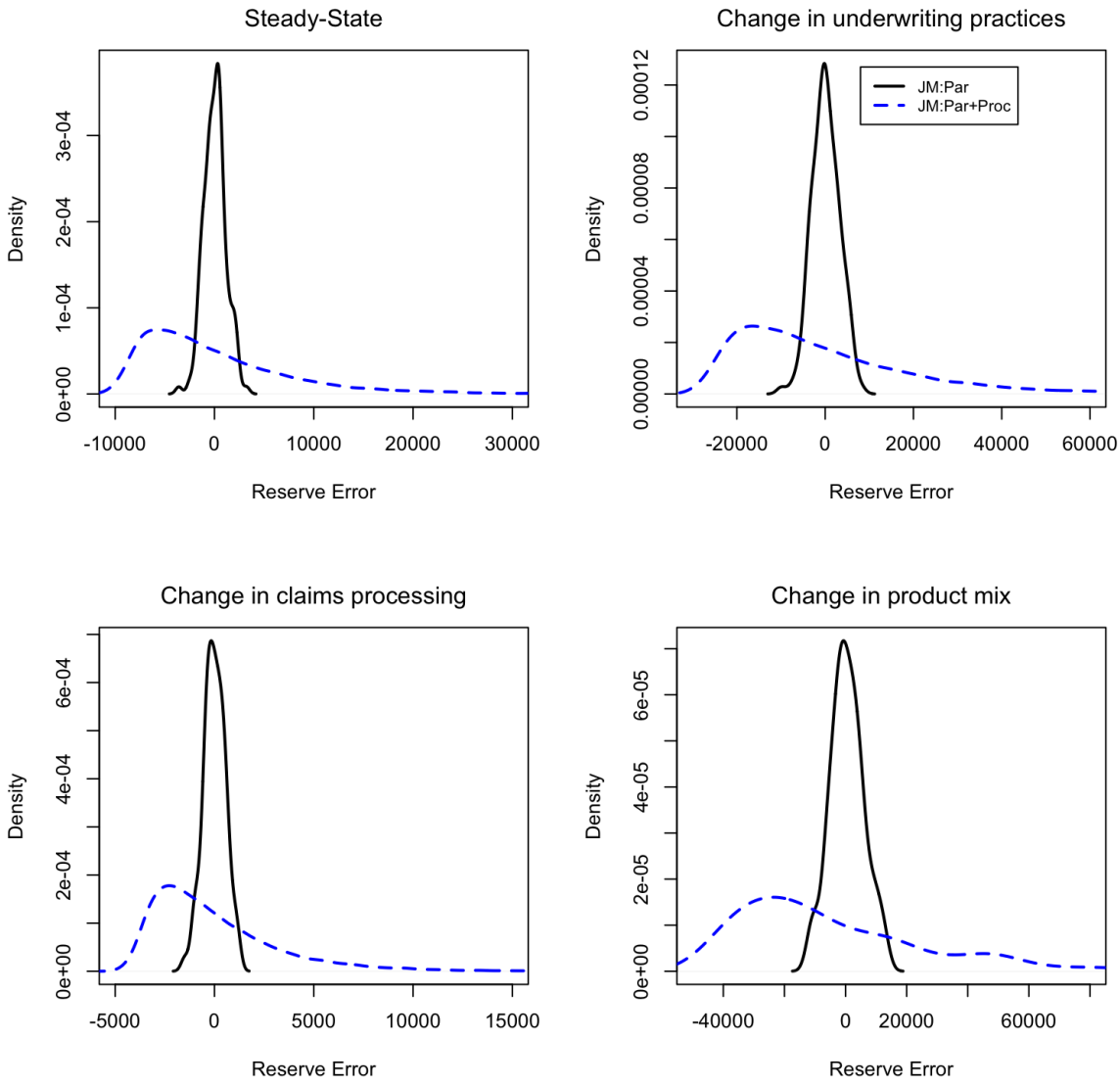

Figure 3.6: Comparing parameter and process uncertainty for the JM under a steady-state and changes in environmental conditions.

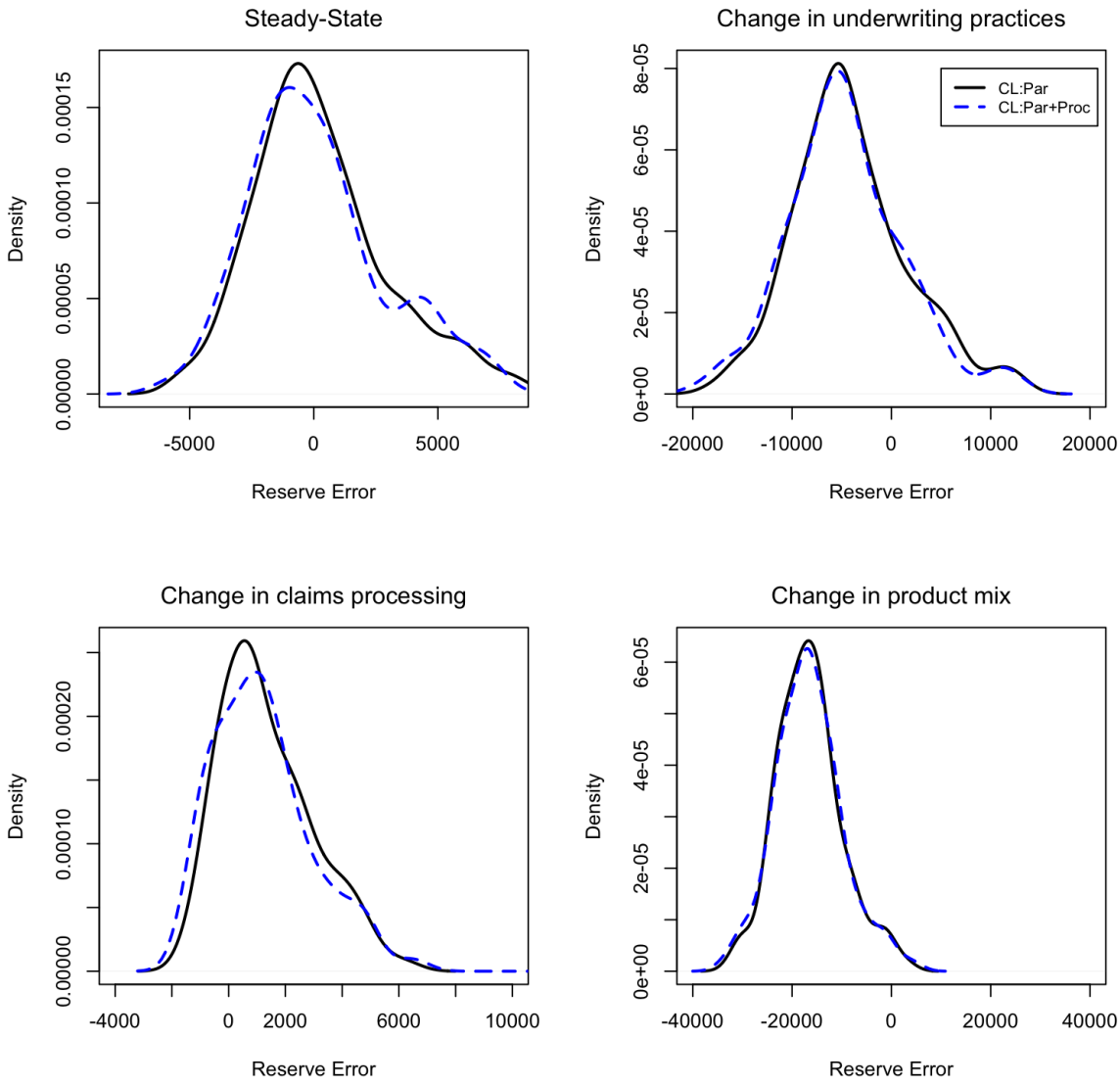

Figure 3.7: Comparing parameter and process uncertainty for the CL method under a steady-state and changes in environmental conditions.

Chapter 4

Empirical Analysis of the Joint Model Framework

Chapter Preview. In this chapter, the joint model (JM) is applied to claims data from a property insurance provider with the focus on RBNS reserve prediction. The joint model is fitted to a training dataset, and I show that accounting for the payment-settlement association helps to accurately predict the settlement time and the ultimate amount of unsettled losses. The RBNS prediction performance of the JM is compared to existing reserving models using an out–of–sample data. Because of the time dimension involved with the RBNS reserve prediction, the traditional cross-validation techniques cannot be used to assess the prediction error of micro-level models. Thus, I introduce a novel form of cross-validation for longitudinal data that I call double cross-validation.

Section 4.1 gives more background information about the data from the Wisconsin Local Government Property Insurance Fund (LGPIF) used in this study. Section 4.2 provides estimation results from the joint model using a training dataset. Section 4.3 evaluates the quality of prediction from the joint model using a hold-out sample and presents a comparison to results from other micro-level techniques and the chain-ladder method. Section 4.4 concludes.

4.1 Data

The data analyzed in this chapter are from the LGPIF, which was established to make property insurance available for local government units. The LGPIF offers three major types of coverage for local government properties: building and contents, inland marine (construction equipment), and motor vehicles. The Fund closed in 2017. When it was operational, on average, it wrote approximately $25 million in premiums and $75 billion in coverage each year; and it insured over a thousand entities.

Exposure information is available from January 1, 2006, to December 31, 2013, and I focus on claims from the building and contents coverage. The training data contain claims that have occurred and were reported between January 1, 2006, and December 31, 2009, as shown in Figure 4.1. 87% of the claims in the training dataset are settled with a single payment. As seen in Table 7.1, there are three different types of payment transactions during the development process of claims in the training data. These are intermediate payments, payments to the settlement of a claim (this can be a single payment to settlement or a payment to settlement after several intermediate payments), and payments after claims are reopened. Other important features of the data are discussed in Section 1.3.1. The data set from January 1, 2010, to December 31, 2013, is the validation dataset. The validation dataset contains the actual unpaid losses used to evaluate the quality of the reserve predictions from the fitted models. The training sample contains 3,393 reported claims including 129 claims reported but with no payment transaction by the valuation date. The total RBNS reserve from the validation sample is $4,511,490 from 163 claims.

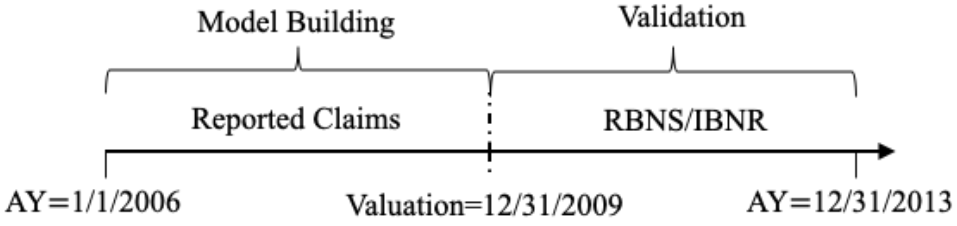

Figure 4.1: Timeline for model fit and prediction.

Table 4.1: Description of variables in the LGPIF data.

Variable	**Description**
	Claim/Transaction Covariates
LnInitialEst	Initial estimates for reported claims in logarithmic of dollars
ReportDelay	Reporting delay
LossYear	Year of claim occurrence
LossQtr	Quarter in the year of claim occurrence
CauseCode	A code to identify the peril type of each claim
TimeToPayment	Time from reporting to payment (Development period)
	Policy/Policyholder Covariate
EntityType	Categorical variable that is one of six types: Village, City, County, Misc, School, or Town
CountyCode	A code to identify which of the 72 counties the entity belongs to
Region	Categorical variable which identifies region the county belongs to: Northern,Northeastern, Southeastern, Southern,or Western
LnPolicyDed	Deductible for the policy in logarithmic of dollars

Table 4.1 describes the covariate information about the policy, policyholder, claim, and transactions used in the model building. From Table 1.1, it is seen that the ultimate claim severity, deductible, and initial estimate distributions are right-skewed. To handle the skewness, I will utilize logarithmic transformations of deductibles and initial estimates.

4.2 Estimation Results

The joint longitudinal-survival framework is applied to the micro-level reserving problem using the property data from the Wisconsin LGPIF. I begin by fitting a base model where for the longitudinal submodel, the observed cumulative payments is assumed to follow a Log-Normal distribution, i.e. $y_{it} \sim \text{Lognormal}(\eta_{it}, \sigma^2)$, and fit a proportional hazard model with a Weibull baseline hazard for the survival submodel. Also, a random intercept longitudinal submodel is assumed where the random effects follow a normal distribution, $\mathcal{N}(0, \nu)$. See Table 7.5 in the Appendix for the estimation results for the base joint model.

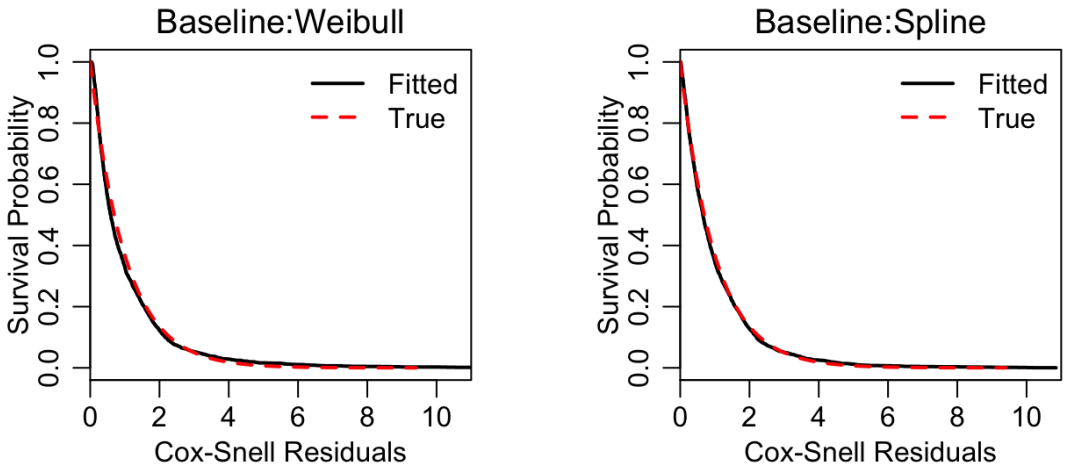

Figure 4.2: Evaluating the goodness of fit for survival submodel.

4.2.1 Evaluation of Survival Submodel Fit

From Section 3.3.2, the correct specification of the survival submodel is necessary to obtain accurate prediction results. In Table 7.5, the baseline hazard function assumes a Weibull model. In this section, the overall survival submodel fit using the Weibull baseline hazard model is compared to a more flexible survival submodel with a spline baseline hazard model. The spline baseline model was fitted with equally-spaced five internal knots in the quantiles of the observed event times.

To examine the survival model's overall fit, I compare the Kaplan-Meier estimate of the Cox-Snell residuals from both survival submodels to the function of the unit exponential distribution graphically (Rizopoulos, 2012). Figure 4.2 plots the fit for the survival submodel with a Weibull and spline baseline hazard functions and assumes a Log-Normal distribution for the longitudinal submodel. The solid line is the Kaplan-Meier estimate of the survival function of the Cox-Snell residuals, and the dashed line is the survival function of the unit exponential distribution. It can be seen that both the Weibull and spline baseline fits the data very well. But I choose the Weibull model because it is easier to interpret its components.

4.2.2 Evaluation of Longitudinal Submodel Fit

For the evaluation of the longitudinal submodel fit in the base model, I first investigate whether the fit of the longitudinal submodel can be improved by assuming a Gamma regression with dispersion parameter $1/\sigma$ and a log-link. The AIC and BIC for the joint model with Log-Normal distribution with linear trend are 74,117 and 74,488, respectively, and that of the Gamma model with linear trend are 73,887 and 74,258, respectively. Therefore, a comparison using AIC and BIC suggests the Gamma model offers a better fit.

I also investigate whether the fit of the longitudinal submodel can be improved by using a non-linear payment trend in the systematic component of the Gamma model. The left panel of Figure 4.3 plots the observed trend overlayed with the fitted linear trend, and the right panel plots the observed trend overlayed with the fitted non-linear trend using B-spline basis function with an internal knot at payment time 5 (payment time in quarters). The AIC and BIC are 73,850 and 74,240 for the non-linear payment trend using splines. The AIC and BIC suggest a slightly better fit with the non-linear trend, and since the correct specification of the payment trend plays a critical role in the prediction of unpaid losses, I choose the model with the non-linear payment trend.

The estimation results for the final fitted joint model, where y_{it} follows a Gamma distribution with a non-linear payment trend and log link in the longitudinal submodel and a Weibull baseline hazard in the survival submodel, are given in Table 4.2. I present the parameter estimates and standard errors of the continuous covariates. For the categorical covariates, I present their likelihood ratio test statistic, the degrees of freedom, and p-value to test the importance of the categorical variable in each submodel. For the survival submodel, the association parameter $\alpha = -0.407$, and it measures the percentage reduction in hazard or risk of the settlement while expected payments increases by one percent. Note that α is highly significant at a 5% significance level and being negative in the hazard model means that the association between the settlement time and payment size is positive.

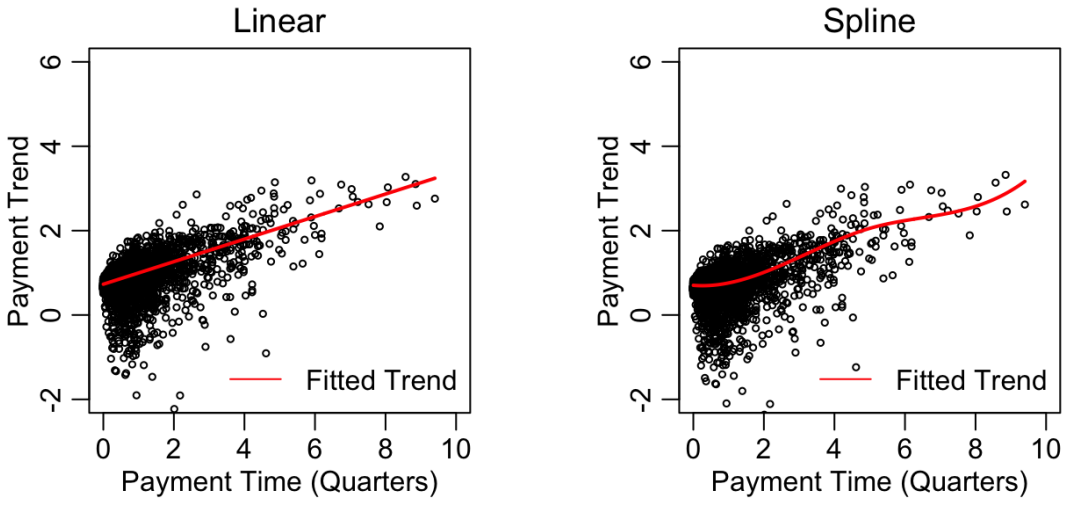

Figure 4.3: Evaluation of payment trend under the longitudinal submodel.

4.3 Out-of-Sample Validation

With the validation data that spans from January 1, 2010, to December 31, 2013, the actual future development trajectory of the RBNS claims after the valuation date can be followed and compared to the predictions from the joint model and other reserving models. In this section, I provide the prediction results for the Independent and Two-Stage estimation techniques. I also provide prediction results from an estimation technique that employs a GLM model for ultimate payments in the payment submodel and a survival submodel that is modeled separately, setting $\alpha = 0$. Further, I present results from the MPP model; after reporting, the transaction occurrence times, the type of transaction, and the transaction's payment amount are considered to be the marks. Different models are specified for each component of the MPP model. In addition, I provide results from the chain-ladder model.

Table 4.2: Estimation results for final joint model: Assuming Gamma distribution with a log link and non-linear payment trend for the longitudinal submodel and a Weibull baseline hazard for the survival submodel.

Longitudinal submodel			Survival submodel		
Variable	Estimate	Std. Error	Variable	Estimate	Std. Error
(Intercept)	0.704	0.108	LnInitialEst	-0.069	0.060
B_1	-0.118	0.089	LnPolicyDed	0.010	0.013
B_2	1.876	0.168	ReportDelay	0.351	0.019
B_3	1.561	0.291			
B_4	2.465	0.343			
LnInitialEst	0.894	0.009			
LnPolicyDed	0.029	0.007			
ReportDelay	-0.012	0.014	α(association)	-0.407	0.067
Variance Components			Weibull Baseline Hazard		
shape (σ)	5.276		λ	50.159	
$\nu^{(1/2)}$	0.417		k	1.459	
Number of Payments	3,891		Number of Claims	3,264	
Categorical Variables					
Variable	LRT	df (p-value)	Variable	LRT	df (p-value)
CauseCode	93.550	9 (<0.0001)	CauseCode	93.430	9 (<0.0001)
Region	24.100	4 (0.0001)	Region	59.860	4 (<0.0001)
EntityType	9.720	5 (0.0837)	EntityType	64.840	5 (<0.0001)
LossQtr	4.120	3 (0.2486)	LossQtr	25.090	3 (0.0001)
LossYear	11.860	3 (0.0079)	LossYear	23.600	3 (<0.0001)

4.3.1 Point Prediction

To get the RBNS reserve estimate from the fitted joint model, I follow the prediction routine in section 2.5.2. Given that the B-splines is been used in the longitudinal submodel, prediction for the ultimate losses is continued linearly for predicted settlement times greater than the largest observed payment times. The Gamma distribution is assumed for the longitudinal submodel for the JM, Independent, the Two-Stage, and the GLM approach. Further, the prediction routine for the Independent, the Two-Stage, and the GLM techniques is similar to that of the joint model. For the MPP, a discrete survival model with piece-wise constant hazard rates is specified for the transaction occurrence, a logit model is specified for the transaction type, and a Gamma regression is specified for the incremental payments. Detailed discussions on the MPP is provided in Appendix 7.4.2. I follow the prediction routine for the RBNS reserve in Antonio

and Plat (2014). The prediction routine simulates the next transaction's exact time, the transaction type (payment to a settlement, or intermediate payment), and the corresponding payment. For the chain-ladder, I employ a modified version of Mack model (Mack, 1993), where claims in the run-off triangle are aggregated using reporting quarter and observation quarter instead of the occurrence quarter and development quarter. Then projections made from these development factors give us RBNS reserve estimates.

Table 4.3 presents the reserve error, which is the expected RBNS reserve minus the actual unpaid losses and the error as a percentage of the actual unpaid losses for JM and other models. For all models except the chain-ladder model, the estimated micro-level model is used to predict the RBNS estimate of each open claim and then aggregated to obtain the reserve estimate for the portfolio. In the out-of-sample data, I consider two claims as "unusual claims" because they had payments totaling over a million dollars at the valuation date. These claims were caused by hail damage to buildings of a school in the year 2007 and a roof collapse of a building in the year 2008 with total payments at the valuation date of $\$5,398,051$ and $\$1,802,742$, respectively. Further, the ultimate amounts of these claims are $\$6,615,117$ and $\$1,842,242$, respectively. At the valuation date, the average total payment of open claims, including the unusual claims, is $\$60,668$, and that of open claims without the unusual claims is $\$16,696$. Naturally, the analyst will remove these unusual claims before any prediction exercise, but as a robustness check, I provide the prediction results with and without the unusual claims.

From the results, JM produced the least percentage reserve error at 0.41% without the unusual claims and a very competitive percentage error of -7.24% with unusual claims. The results from the MPP are also competitive compared to JM. The performance of the Two-Stage and Independent estimation techniques, in comparison to the JM, emphasizes that when the association between the payment process and settlement process and the endogenous nature of the payments process is ignored, it leads to inaccurate prediction of unpaid losses. Without any surprise, the GLM estimation technique, which only utilizes the ultimate payment and ignores the payment-settlement association, did not

perform well. The results also show the chain-ladder method did not perform well in estimating the unpaid losses with the unusual claims but was very competitive without the unusual claims.

Table 4.3: RBNS reserve point prediction results for the validation sample.

	Without unusual claims		With unusual claims	
	RBNS Estimate	Error %	RBNS Estimate	Error %
True Reserve	3,254,924		4,511,490	
JM Error	13,505	0.41	-326,721	-7.24
MPP Error	450,747	13.85	57,901	1.28
Two-Stage Error	74,910	2.30	451,714	10.01
Independent (JM with $\alpha = 0$) Error	725,206	22.28	-506,418	-11.23
GLM (Closed claims) Error	1,937,817	59.53	3,187,220	70.65
Chain-Ladder Error	262,232	8.06	1,261,332	27.96

The left panel of Figure 4.4 shows the comparison of the distribution of actual ultimate losses and predicted ultimate losses from JM over time. It can be seen that JM provides accurate predictions over time. Another advantage of the joint model is that it can be used to predict the time to settlement for open claims, which will be particularly useful in the run-off operation of an insurer. For example, in a run-off situation for a workers compensation insurer, losses for which claimants would not take an offered settlement usually involves regular payments until death (Kahn, 2002). Therefore accurately predicting the settlement time or remaining months to live is important in the reserving exercise. The right panel of Figure 4.4 provides a comparison of actual settlement times and predicted settlement times using the joint model. The joint model accurately predicts the settlement times with a Spearman correlation coefficient of 83%.

4.3.2 Predictive Distribution

Here, I am not only interested in the expected value of prediction but also the variability in prediction. As a measure of reserve uncertainty, I provide the standard error, which is the standard deviation of the predictive distribution accounting for only parameter uncertainty and the root mean squared error of prediction (RMSEP), which is the standard deviation of the predictive distribution after accounting for both parameter

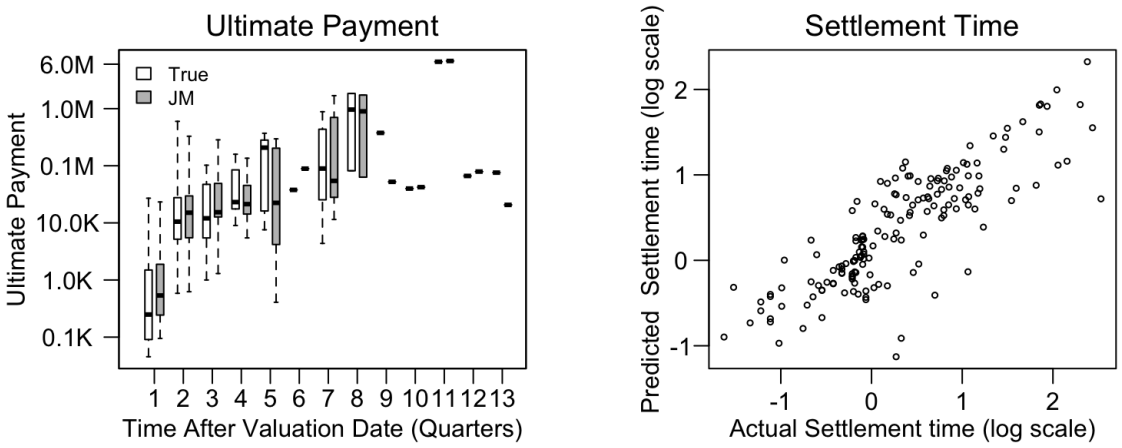

Figure 4.4: Left Panel: Distribution of the true and predicted ultimate payment over time (with unusual claims). Right Panel: Comparison of actual settlement times and predicted settlement times using JM (with unusual claims).

and process uncertainty (England and Verrall, 2002). All the prediction results in this section are based on $10,000$ replications.

4.3.2.1 Predictive Distribution of the Expected Unpaid Losses

The predictive distribution of the expected outstanding payments is obtained by incorporating only the parameter uncertainty. For the joint model, I assume that the parameter estimates can be approximated by a multivariate normal distribution with the maximum likelihood estimates $\hat{\theta}$ as mean and covariance matrix $\widehat{Var}(\hat{\theta})$. The routine for the distribution of the expected outstanding payments is elaborated in Algorithm 3. The total RBNS liability for each replication is obtained by adding the RBNS prediction for all claims.

The predictive distribution routine for the Independent, the Two-Stage, and the GLM estimation techniques follow a similar procedure as the JM. For the MPP, I repeat the prediction routine to predict RBNS reserve in (Antonio and Plat, 2014), and I take the expected values for the payment amounts model. To obtain the predictive distribution of

the expected values of unpaid losses for the CL, I employ the bootstrapping algorithm in England and Verrall (2002) and implemented in the `ChainLadder` R package (Carrato et al., 2020).

Table 4.4 presents the standard error for the out-of-sample data without the unusual claims. Not surprisingly, the joint model produced a significantly lower standard error than that of the chain-ladder. The higher predictive uncertainty of the chain-ladder is due to the loss of information from data aggregation. Also, building the reserving model with only information from closed claims, as seen with the GLM method, leads to a higher predictive uncertainty as well. The standard error from the MPP is higher than the joint model because the MPP model is composed of three submodels containing more parameters than two submodels from the joint model. By accurately accounting for the payment-settlement association, the joint model produced slightly higher standard error compared to the Independent and Two-Stage techniques. Figure 4.5 presents an illustration of the predictive distribution of the expected reserve estimates focusing on the out-of-sample data without unusual claims, and it can be seen that the JM provides both accurate mean prediction and low predictive uncertainty.

Table 4.4: RBNS reserve predictive distribution results for the validation sample (without unusual claims).

	Estimate	SE	RMSEP
True Reserve	3,254,924		
JM	3,268,429	430,847	2,078,825
MPP	3,705,671	630,956	1,381,430
Two-Stage	3,329,834	336,761	2,084,256
Independent (JM with $\alpha = 0$)	3,980,130	343,737	2,261,560
GLM (Closed claims)	5,192,741	791,364	4,849,763
Chain-Ladder	3,517,156	987,054	1,334,597

4.3.2.2 Predictive Distribution of Losses

For the predictive distribution of losses, in addition to the parameter uncertainty, the process uncertainty is introduced to match the randomness of the development of losses. The ultimate payments are generated using the process distribution in the longitudinal

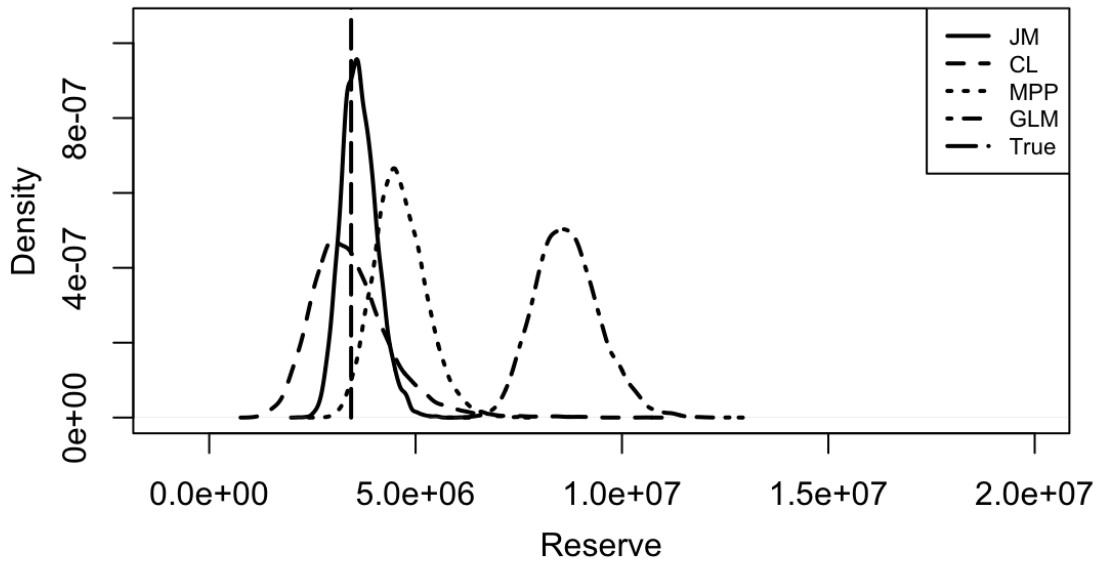

Figure 4.5: Predictive distribution of expected reserve estimates (without unusual claims).

submodel at each replication. I repeat the steps in Algorithm 3 for introducing parameter uncertainty and introduce process uncertainty by simulating the ultimate payments from the process distribution of the longitudinal submodel given each simulated set of parameters. The RBNS liability is then calculated for each simulated ultimate loss, and the total RBNS liability for each replication is obtained by adding the RBNS prediction for all claims. Again, the predictive distribution for the Independent, Two-Stage, and GLM techniques follow a similar procedure as JM. For the MPP, the process uncertainty is introduced by simulating payments from Gamma distribution. I account for the process uncertainty in the CL method by simulating payments in the future cells in the run-off triangle from the over-dispersed Poisson (England and Verrall, 2002).

Table 4.4 also presents the RMSEP for the out-of-sample data without the unusual claims, and Figure 4.6 shows the predictive distribution after accounting for both parameter and the process uncertainty from the JM and other models. It can be seen that the joint model is associated with a higher process variance hence higher RMSEP compared to the MPP and the chain-ladder. As discussed in Section 3.5, the process variance from the joint model is higher because it is implemented using cumulative payments in the longitudinal submodel. The MPP and the chain-ladder are implemented on the incremental payments.

4.3.3 Double Cross-Validation

In this subsection, I quantify the prediction error of different individual reserving methods using a novel out-of-sample validation method, which I call double-cross validation. The novelty of this approach comes from the longitudinal nature of the claims payment process, which makes it impossible to utilize traditional cross-validation techniques. Here, on the time dimension, I split the data by the valuation date. On the cross-section dimension, I split the data by the reporting date. So the training data contains payment from claims that have been reported by the valuation date. Then, the out-of-sample data comprises two parts. The first part contains payments made after the valuation

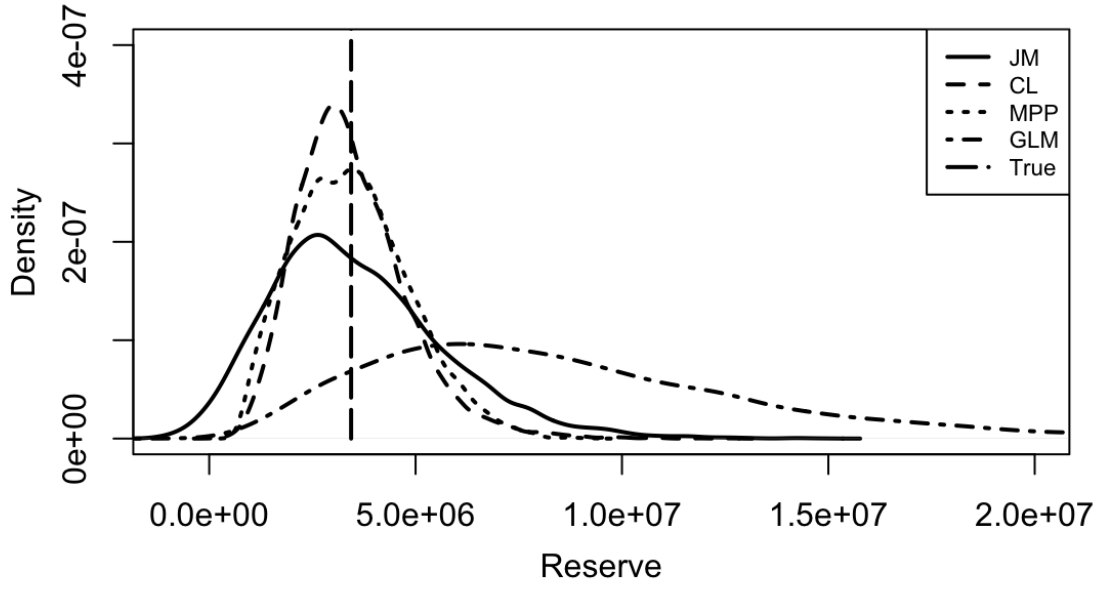

Figure 4.6: Predictive distributions (Parameter + Process Uncertainty) of the total RBNS reserve (without unusual claims).

date on claims reported before the valuation date; I call this the validation dataset. The second part contains payments from newly reported claims during the out-of-sample period, and I call this the test dataset. The routine for a K-fold double cross-validation technique is outlined in Algorithm 4. See Figure 4.7 for an example of 10-fold double cross-validation.

The prediction error percentages are obtained from both the validation and test datasets, and Table 4.5 provides the mean percentage error from the 10-fold double cross-validation for the JM and other micro-level models. The results for the CL method is not provided because the cross-validation technique cannot be applied to macro-level models. Overall, the mean prediction percentage error for JM in the validation and test datasets are better than the results from other models, which emphasizes the robustness of the model.

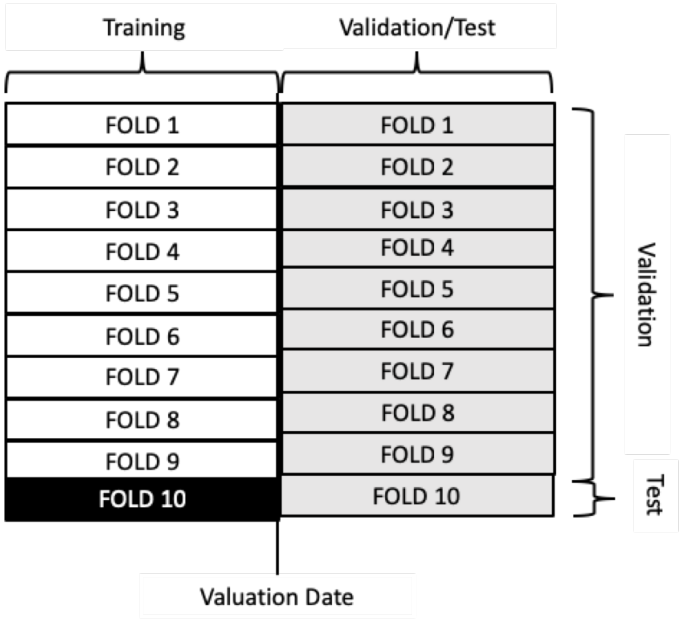

Figure 4.7: 10-fold double cross-validation technique.

4.3.4 Discussion on IBNR Reserving

This chapter focuses on RBNS claims, so the practicing actuary would need to combine the joint model approach with a method for estimating IBNR claims to obtain the IBNR

Algorithm 3 Reserve predictive distribution for JM using Monte Carlo simulation.

Input: Valuation time c_i, observed data at valuation $(t_i, \delta_i, \mathbf{y}_i, \mathbf{w_{ic_i}}, \mathbf{x_{ic_i}}, \mathbf{z_{ic_i}})$, covariates at future time u $(\mathbf{w_{iu}}, \mathbf{x_{iu}}, \mathbf{z_{iu}})$, ML estimates $\hat{\theta}$, $Var(\hat{\theta})$, cumulative amount paid $Y_i(c_i)$, and empirical Bayes estimate $\hat{\mathbf{b}}_i$, number of draws K, and number of replications L.

Output: $\left\{\hat{R}_i^{RBNS_l}(c_i), l = 1, ..., L\right\}$;

for $l = 1, ..., L$ **do**
- Generate $\theta^l \sim \mathcal{N}(\hat{\theta}, Var(\hat{\theta}))$;
- Generate $\mathbf{b}_i^l \sim f(\mathbf{b}_i | t_i, \delta_i, \mathbf{y}_i; \theta^{\mathbf{l}})$;
- Calculate $S_i^l(c_i) = \exp\left(-\int_0^{c_i} h_0^l(s) \exp\{\gamma'^l \mathbf{w}_{is} + \alpha^l \eta_{is}^l\} ds\right)$; where $\eta_{is}^l = \mathbf{x}'_{is}\beta^{\mathbf{l}} + \mathbf{z}'_{is}\mathbf{b}^{\mathbf{l}}{}_i$ and $\{\alpha^l, \gamma^{\mathbf{l}}, \beta^{\mathbf{l}}\} \in \theta^l$;
- **for** $k = 1, ..., K$ **do**
  - Generate $\hat{\pi}_i(u|c_i) = U_k \sim \text{Uniform}(0, 1)$;
  - Calculate $u_{ik}^l = H_i^{-1}(-\log(U_k \times S_i^l(c_i)))$; where $H_i(u) = \int_0^u h_0^l(s) \exp\{\gamma'^l \mathbf{w}_{is} + \alpha^l \eta_{is}^l\} ds$;
- **end for**
- **return** $\left\{u_{ik}^l; k = 1, ..., K\right\}$;
- Calculate $\hat{u}_i^l = K^{-1} \sum_{k=1}^K u_{ik}^l$;
- Generate $\hat{Y}_i^{ULT_l}(\hat{u}_i^l) = \exp(\hat{\eta}_{i\hat{u}_i^l})$; For parameter uncertainty.
- Generate $\hat{Y}_i^{ULT_l}(\hat{u}_i^l) \sim \text{Gamma}\left(\frac{\exp(\hat{\eta}_{i\hat{u}_i^l})}{\sigma^l}, \sigma^l\right)$; For parameter and process uncertainty. where $\hat{\eta}_{i\hat{u}_i^l} = \mathbf{x}'_{i\hat{u}_i^l}\beta^l + \mathbf{z}'_{i\hat{u}_i^l}\mathbf{b}_i^l$ and $\{\beta^l, \sigma^l\} \in \theta^l$;
- Calculate $\hat{R}_i^{RBNS_l}(c_i) = \hat{Y}_i^{ULT_l}(\hat{u}_i) - Y_i(c_i)$;
- **return** $\left\{\hat{R}_i^{RBNS_l}(c_i), l = 1, ..., L\right\}$;

end for

Algorithm 4 Double cross-validation technique.

Input: Valuation time c and full dataset $\mathcal{D}_t = \{\mathcal{D}_t^T, \mathcal{D}_t^V\}$; where $\mathcal{D}_t^T$ is the training dataset, $\mathcal{D}_t^V$ is the validation dataset and t represent claim payment times;

Output: $\left\{\psi_k^V, k = 1, ..., K\right\}$ and $\left\{\psi_k^T, k = 1, ..., K\right\}$;

Split $\mathcal{D}_t$ into K groups, $\mathcal{D}_t = \{\mathcal{D}_t^k\}_{k=1}^K$;

for $k = 1, ..., K$ **do**
- Generate model building dataset $\mathcal{D}_t^S = \{\mathcal{D}_t^k\}_{k \neq k}$;
- Generate hold-out dataset $\mathcal{D}_t^H = \{\mathcal{D}_t^k\}_{k=k}$;
- Fit the model using training dataset $\mathcal{D}_t^{S_T} = \{\mathcal{D}_t^S\}_{t \leq c}$;
- Generate validation dataset $\mathcal{D}_t^{S_V} = \{\mathcal{D}_t^S\}_{t > c}$;
- Generate test dataset $\mathcal{D}_t^{H_T} = \{\mathcal{D}_t^H\}_{t > c}$;
- Calculate prediction error percentage ψ_k^V using $\mathcal{D}_t^{S_V}$;
- Calculate prediction error percentage ψ_k^T using $\mathcal{D}_t^{H_T}$;
- **return** $\left\{\psi_k^V, k = 1, ..., K\right\}$ and $\left\{\psi_k^T, k = 1, ..., K\right\}$;

end for

Table 4.5: Mean percentage error from 10-fold double cross validation (without unusual claims).

	Validation data error %	Test data error %
JM	3.99	-17.74
MPP	47.72	-12.26
Two-Stage	26.59	-44.92
Independent (JM with $\alpha = 0$)	47.66	-25.25
GLM (Closed Claims)	83.92	87.32

reserves. The general framework for estimating IBNR reserves can be broken down into two stages. The first stage involves modeling the number of IBNR claims and their reporting delays with chain-ladder type strategies; for example, see Martínez-Miranda et al. (2012) and Wüthrich (2018a). Further, Crevecoeur et al. (2019) proposes a granular approach to model the number of IBNR claims due to the heterogeneity of the reporting delay based on claim occurrence day and calendar day effects such as weekday and holiday effects. The second stage involves modeling the development of the predicted IBNR claims with the proposed joint model fitted using accident date and reporting delay as covariates.

4.4 Conclusion

In claims management, the settlement duration is usually positively associated with the size of the claim. The payment-settlement association means settlement times will be impacted by paid losses, which affects the reserve prediction of open claims. Therefore, ignoring the payment-settlement association could lead to inaccurate predictions of outstanding payments.

In this chapter, to incorporate the correlation between the payment and the settlement processes, the joint longitudinal-survival model (JM) framework was applied to the reserving problem using data from a property insurance provider. The prediction results from the joint model is compared to existing reserving models, and the results show that accounting for the payment-settlement association leads to better prediction accuracy

and lower reserve uncertainty compared to models that ignore it.

I also introduced a novel cross-validation technique named double cross-validation as a result of the time dimension involved with claim development. The double cross-validation technique provides two datasets (validation and test datasets) for the evaluation of the robustness of the models. The validation dataset contains outstanding payments for claims reported by the valuation date. The test data contains payments from newly reported claims during the out-of-sample period. Again, the joint model displayed superior prediction accuracy using both datasets compared to models that ignore the payment-settlement association, which highlights the robustness of the model.

Chapter 5

Improving Ratemaking Using Micro-Level Loss Prediction Techniques

Chapter Preview. In pricing insurance contracts for non-life insurers, the literature has mainly focused on using detailed information from policies and closed claims. Information regarding RBNS and IBNR are usually ignored during ratemaking. This chapter employs a micro-level reserving technique to incorporate open claims in insurance pricing.

Section 5.1 introduces the problem, and Section 5.2 provides information about the ratemaking data that inspires the proposed modeling framework. Section 5.3 presents the marked Poisson process model and its application to ratemaking. Section 5.4 provides the model fitting results from the marked Poisson process model using a training dataset. Section 5.5 evaluates the quality of prediction from the marked Poisson process model using out-of-sample data. Section 5.6 concludes.

5.1 Introduction

In non-life insurance, premiums are set to cover the expected future cost and also allow for the earmarked underwriting profit through a process known as ratemaking. The pure premium method and the loss ratio method are the traditional techniques for ratemaking, and these techniques are focused on whether the total premiums will cover the total costs. See Werner and Modlin (2016) for details on these techniques. Accurate risk pricing is expected to provide stronger incentives for more caution, resulting in lower claim frequencies and reductions in insurance loss costs (Cummins, 2002). To better align premiums with expected costs, actuaries develop rates by employing multivariate risk classification techniques based on information from the policy and the claim history. The multivariate risk classification techniques make it possible to account for several risk factors simultaneously. The generalized linear models (GLMs) and machine learning algorithms are two popular techniques for multivariate risk classification (Werner and Modlin, 2016).

The frequency-severity model and the Tweedie GLM are two popular approaches used to model the claim frequency and payments arising from the closed claims. See Frees (2014) for details and application of these models. An observation from the ratemaking literature is that the data used in the multivariate analysis is often based on closed claims, where the ultimate amount paid for all claims is known. This observation is not surprising as there is a natural friction between using only closed claims from older policy years and using the information on all reported claims, which will include information on open claims. With closed claims, all uncertainties in information on open claims are eliminated. In contrast, the information on open claims can reflect shifts in the distribution of the expected claim payments better than closed claims. Therefore, ignoring open claims during the ratemaking process may lead to biased estimates and, consequently, inaccurate premiums. Practicing actuaries are well aware of these biases, so they make ad hoc adjustments to address them. But, by their very nature, these adjustments depend on the ability of the actuary. The profession is better served by

having formal procedures to make adjustments.

This chapter presents an intuitive framework for ratemaking that ensures that the multivariate risk analysis is done using the information on claims that have been closed, and payments on open claims. To model the complete development of claims for ratemaking purposes, I employ the marked Poisson process (MPP) framework with four hierarchical building blocks. Three of the building blocks drive the expected cost based on reported claims by modeling the number of claims per policy in a policy year, the conditional number of payment transactions for a claim, and the conditional payment sizes for each transaction. One advantage of the MPP is that the likelihood of the claims process can be decomposed into independent blocks, which allows each block to be maximized in isolation (Larsen, 2007). As a result, the parameters of each block are estimated with the appropriate GLMs. For RBNS claims, the number of transactions is censored at the ratemaking date, which is duly addressed. I use policy covariates that are readily available for new and existing policyholders for an observation period. The fourth building block accounts for the expected future cost relating to IBNR claims through a unique feature of the MPP framework by analyzing the reporting delay distribution of claims.

As discussed in Section 1.2.2, the MPP framework, which was introduced by Arjas (1989), Jewell (1989), Norberg (1993), and Norberg (1999) has been widely used for individual-level loss reserving. For example, Antonio and Plat (2014) and Verrall and Wüthrich (2016) apply the marked Poisson process for non-life insurance loss reserving where claims occurrences are assumed to follow a non-homogeneous Poisson process, and stochastic characteristics about the claims are treated as marks. The MPP framework's hierarchical makeup also provides flexibility in modeling different events and their features in the ratemaking process. In insurance pricing, hierarchical models are not new. For example, the frequency-severity model forms a two-level hierarchical pricing model. Frees and Valdez (2008) and Frees et al. (2009) extended the frequency-severity model to a hierarchical model with three building blocks relating to the frequency, type, and severity of claims. Shi et al. (2016) provides a hierarchical framework for

modeling insurance loss cost with a complex structure and proposes a copula regression to accommodate various sources of dependence.

5.2 Empirical Motivation

The data I use for the ratemaking exercise in this chapter is from the Wisconsin Local Government Property Insurance Fund (LGPIF) described in Section 4.1. Though the LGPIF data spans from January 1, 2006, to December 31, 2013, I focus on the dataset from effective years 2006-2011, where all claims are marked as closed as of December 31, 2013. Here, I use data from the policy, claim, and transaction databases. Table 5.1 shows the summary statistics, at the policy and claim level, from effective years 2006-2011. High variability across years in the average claim frequencies and severity is observed at the policy level, highlighting the importance of using current information in claim modeling for ratemking purposes. Further, from the summary statistics at the claim level, the average number of payments transactions to settlement per claim is gradually reducing, and the average payment per transaction is increasing. This observation suggests a change in claims processing of the LGPIF. As discussed in Section 3.4, environmental changes affect the distribution of future losses, and using current information on open claims allows to capture such changes promptly.

Table 5.1: Summary statistics at the policy and claim level for building and contents coverage.

	Policy Level				Claim Level	
Effective Year	Average Frequency	Average Severity	Average Coverage (Million)	Number of Policies	Average No. of Transaction	Average Payment Per Transaction
2006	0.734	10,083	32.363	1,159	1.276	9,554
2007	0.925	7,095	35.143	1,143	1.291	8,946
2008	0.746	6,730	37.150	1,130	1.245	7,991
2009	0.924	4,864	40.275	1,114	1.206	9,864
2010	1.088	20,827	41.123	1,114	1.123	17,152
2011	0.948	8,367	42.426	1,096	1.173	17,156

Table 5.2 describes the rating variables considered in this chapter. Tables 7.6 and 7.7 in the Appendix 7.5 show that the rating variables are correlated with the claim frequency

and severity at the policy level, and with the transaction frequency and severity at the claim level, which indicates that they will be significant predictors to claims in the ratemaking model.

Table 5.2: Description of rating variables.

Variable	**Description**
EntityType	Categorical variable that is one of six types: Village, City, County, Misc, School, or Town
Region	Categorical variable which identifies region the county of an entity belongs to: Northern,Northeastern, Southeastern, Southern,or Western
LnPolicyDed	Deductible for the policy in logarithmic of dollars
LnPolicyCov	Total building and contents coverage for the policy in logarithmic millions of dollars
AlarmCredit	Categorical variable that is one of five types:(0%, 5%, 10%,15%, or a combination of credits), for automatic smoke alarms in main rooms

For the ratemaking exercise, the data from effective years 2006-2009 are used as the training sample to calibrate the MPP model. Here, I assume that by December 31, 2009, policyholders' rates have to be updated for the policy year 2010. The rating factors from the calibrated MPP model are then applied to the 2010 rating variables to predict 2010 claim scores. Table 5.3 provides a summary of the number of closed, RBNS, and IBNR claims as of December 31, 2009. As expected, the very recent effective year 2009 is associated with the highest RBNS and IBNR claims. Therefore, ratemaking models that rely on only closed claims will lose current information needed to produce accurate premiums. The table also summarizes the average payments for reported claims (closed and RBNS claims) as of December 31, 2009. It is seen that the average payments across years for RBNS claims are more significant than closed claims indicating that bigger claims take a longer time to settle. Hence, the payments on the RBNS claims can reflect the change in the risk profile of the Fund in a timely manner compared to closed claims.

For robustness checks, I also present an analysis of the MPP framework using the effective years 2007-2010 as the training sample. I compare the claims scores from this new model to the out-of-sample data from 2011.

Table 5.3: Summary statistics for closed, RBNS, and IBNR claims as of December 31, 2009.

	Number of claims				Average payments		
Effective Year	Closed	RBNS	IBNR	Total	Closed	RBNS	Total
2006	785	1	0	786	16,501	5,398,051	23,348
2007	987	6	4	997	13,281	376,109	15,473
2008	747	20	10	777	13,509	51,165	14,491
2009	478	136	184	798	5,631	8,904	6,356
Total	3,230	163	198	3,591	14,133	60,668	16,368

5.3 Claim Modeling

5.3.1 Marked Poisson Process

Figure 5.1 elaborates on the timeline for claim occurrence at times $V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n$ and transaction occurrence at times $S_1 = s_1, S_2 = s_2, \ldots, S_m = s_m$ in a fixed period $[0, \tau]$. In this chapter, τ represents the ratemaking date. From the figure, it is clear there are two counting processes, one relating to claims occurrence and the other the transaction occurrence after reporting.

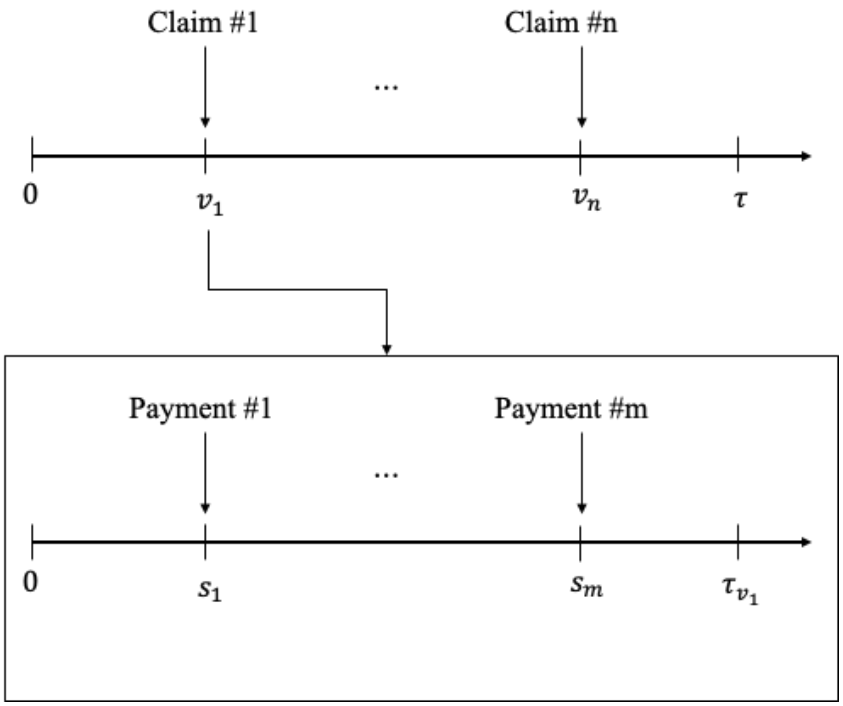

Figure 5.1: Claim occurrence and payment development process.

As described in Section 1.2.2, the associated counting process $\{N(v), 0 \leq v\}$ of the claim occurrence process in Figure 5.1 is Poisson and records the cumulative number of claims that the process generates. I denote $H(v) = \{N(u) : 0 \leq u < v\}$ to be the history of the claims occurrence process at time v. Then the intensity function, determined only by v, for the claim occurrence process as defined in (1.3) is

$$\rho(v|H(v)) = \lim_{\Delta v \downarrow 0} \frac{\Pr\{\Delta N(v) = 1|H(v)\}}{\Delta v} = \lim_{\Delta v \downarrow 0} \frac{\Pr\{\Delta N(v) = 1\}}{\Delta v} = \rho(v). \tag{5.1}$$

Further, observable covariates $x(v)$ that affects claim occurrence may be incorporated in the model by including the covariate information in the process history. Thus, the heterogeneities among the policyholders can be accounted for by specifying the intensity function of the form:

$$\rho(v|x^v) = \rho_0(v) \exp(x'(v)\beta), \tag{5.2}$$

where $x^v = \{x(u) : 0 \leq u \leq v\}$ is the covariate history. $\rho_0(v)$ is the baseline function that relates to policyholders for whom $x(v) = 0$ for all v, and β is a vector of regression coefficients for the covariates.

The marked Poisson process (MPP) framework discussed in Section 1.2.2 is employed for the claims modeling for insurance pricing. Stating the main points again, for a marked Poisson process in $[0, \tau]$, the likelihood that n claims occur at times $V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n$, with marks $Z_1 = z_1, Z_2 = z_2, \ldots, Z_n = z_n$ is given by:

$$\Pr[N = n, (V_i, Z_i)) = (v_i, z_i), i = 1, 2, \ldots, n] = \left(\prod_{i=1}^{n} \rho(v_i) P_{Z|v_i}(z_i) \right) \exp\left(- \int_0^\tau \rho(v) dv \right). \tag{5.3}$$

Here, the claim occurrence counting process $N(v)$ is a Poisson process with intensity function $\rho(v)$. The distribution of the marks $P_{Z|v}$ is conditional on $\Delta N(v) = 1$. After

claim occurrence, the marks can be further broken down into the reporting delay U_i and the claim development process after reporting W_i, i.e., $Z_i = (U_i, W_i)$. As seen in Figure 5.1, W_i includes payment transactions occurrence times S_{ik} and the severity of each transaction P_{ik}. Where $k = 1, \ldots, m_i$ index payment transactions for the ith claim. Then the distribution of the marks $P_{Z|v}$ is specified as $P_{Z|v} = P_{U|v} \times P_{W|v,u}$.

The reporting delay distribution U given occurrence time v, $P_{U|v}$, can be modeled using various distributions from survival analysis, but I specify a mixed distribution comprising of a discrete distribution for a reporting delay below or equal to r days, and a Weibull distribution for reporting delays above r days with density function f_U. The likelihood for the reporting delay is given by:

$$\sum_{r=0}^{d} q_r 1(U = r) + (1 - \sum_{r=0}^{d} q_r) f_{U|U>r}(u), \tag{5.4}$$

where the probability mass for a reporting delay of r days is given by q_r. Specifically, I use $d = 0$, i.e., a probability mass for a reporting delay of zero days (reporting in the same day of occurrence, q_0). To incorporate the policyholder characteristics x_j that may impact the reporting delay distribution of claim i, I specify a Weibull distribution with the scale parameter θ_i that depends on the policyholder characteristics and a constant shape parameter κ given by:

$$f_{U|U>r}(u_i; \kappa, \theta) \sim \frac{\kappa u_i^{\kappa-1}}{{\theta_i}^{\kappa}} \exp[-(u_i/\theta_i)^{\kappa}], \quad \theta_i = \exp(x_j'\gamma). \tag{5.5}$$

The other component of $P_{Z|v}$ is the distribution of the claim development process after reporting, $P_{W|v,u}$. I assume the occurrence of transactions for claim i also follow a non-homogeneous Poisson process in $[0, \tau]$, and transaction payment amounts are treated as marks. Then the likelihood that m_i transactions occur at times $S_{i1} = s_{i1}, S_{i2} = s_{i2}, \ldots, S_{im_i} = s_{im_i}$, with marks $P_{i1} = p_{i1}, P_{i2} = p_{i2}, \ldots, P_{im_i} = p_{im_i}$ is given by:

$$\Pr[M_i = m_i, (S_{ik}, P_{ik})) = (s_{ik}, p_{ik}), k = 1, 2, \ldots, m_i] = \begin{array}{l} (\prod_{k=1}^{m_i} \lambda_i(s_{ik}) f_P(p_{ik})) \\ \times \exp\left(-\int_0^\tau \lambda_i(s) ds\right). \end{array} \tag{5.6}$$

Here, the transaction occurrence counting process $M_i(s)$ is a Poisson process with intensity function $\lambda_i(s)$, and k applies to all payments in $[0, \tau]$. $f_P(p_{ik})$ denotes the density function for the payment severity.

5.3.2 Estimating Parameters

5.3.2.1 Data Structure

Let $j = 1, \ldots, J$ represent the index for policies in the portfolio, and $t = 1, \ldots, T_j$ represent the policy years observed for each policy, then the observable responses at the ratemaking date are:

- N_{jt}, the number of claims reported within a policy year.
- $M_{jt,i}$, the number of transactions for each claim, where $i = 1, \ldots, N_{jt}$ is the claim index. For open claims, $M_{jt,i}$ is censored; then denote $\delta_{jt,i} = 1$ when claim is closed or $\delta_{jt,i} = 0$ otherwise.
- $P_{jt,ik}$, the payment amount per transaction. Where the payment transaction index is $k = 1, \ldots, M_{jt,i}$.

The exposure e_{jt} is measured as a fraction of years, which provides the length of time in the policy year as of the ratemaking date. For the explanatory covariates, policy-level characteristics represented by x_{jt} and described in Table 5.2 are used. Additionally, $U_{jt,i}$ is the reporting delay variable (the difference between claim occurrence and reporting times) for claim i reported in the $\{jt\}$ observation period. Then the data available can be summarized as:

$$\{e_{jt}, N_{jt}, U_{jt,i}, (M_{jt,i}, \delta_{jt,i}), P_{jt,ik}, x_{jt}; t = 1, \ldots, T_j, j = 1, \ldots, J\}. \tag{5.7}$$

5.3.2.2 Reported Claims Modeling

At the ratemaking time τ, as shown in Table 5.3, there are reported claims whose full or partial development process is observed i.e. $\mathcal{C}^{rep} = ((v, u, w) \in \mathcal{C} | v + u \leq \tau)$, and IBNR claims whose development process is totally unobserved i.e. $\mathcal{C}^{ibnr} = ((v, u, w) \in \mathcal{C} | v \leq \tau, v + u > \tau)$. Then, the occurrence of reported claims follows an independent Poisson process with intensity function $\rho(v)F_{U|v}(\tau - v)$ and that of IBNR claims also follows an independent Poisson process with intensity function $\rho(v)(1 - F_{U|v}(\tau - v))$ (Wüthrich and Merz, 2008).

It follows that the observed likelihood of the claims process is given by:

$$L = \left(\prod_{i: v_i + u_i \leq \tau} \rho(v_i) F_{U|v}(\tau - v_i) \right) \exp\left(- \int_0^\tau \rho(v) F_{U|v}(\tau - v) dv \right) \times f_{W|v,u}^{\tau - v_i - u_i}(w_i), \tag{5.8}$$

where $f(\cdot)$ and $F(\cdot)$ denotes a pdf and a cdf, respectively. The superscript in the claim development term (last term in (5.8)) represents that a claim that occurred at v_i and with reporting delay u_i is censored at $\tau - v_i - u_i$ time units after reporting. As discussed earlier, given the occurrence time v and the reporting delay u, the claim development process W can be decomposed into the payment transactions occurrence times S and the severity of each transaction P. Then, the observed likelihood in (5.8) becomes:

$$\begin{aligned} L \;&= \left(\textstyle\prod_{i: v_i + u_i \leq \tau} \rho(v_i) F_{U|v}(\tau - v_i) \right) \exp\left(- \int_0^\tau \rho(v) F_{U|v}(\tau - v) dv \right) \\ &\times \textstyle\prod_{i: v_i + u_i \leq \tau} \left(\prod_k \lambda_i(s_{ik}) \exp\left(- \int_0^{\tau_i} \lambda_i(s) ds \right) \right) \\ &\times \textstyle\prod_{i: v_i + u_i \leq \tau} \prod_k f_P(p_{ik}). \end{aligned} \tag{5.9}$$

Here, k applies to all payments in $[0, \tau_i]$, where $\tau_i = \min(\tau - v_i - u_i, S_i)$. S_i is the total

waiting time from reporting to settlement of claim i. I emphasize that in addition to the claim occurrence counting process $N(v)$ with intensity function $\rho(v)$, the transaction occurrence counting process $M_i(s)$ is also a Poisson process with intensity function $\lambda_i(s)$.

Considering the ratemaking data is organized by the $\{jt\}$ observation period, where $j = 1, \ldots, J$ index policyholders, and $t = 1, \ldots, T_j$ index the policy years for each claim; let the intensity function for the counting process of claim occurrence $N_{jt}(v)$ be $\rho_{jt}(v)$ and that of counting process of transaction occurrence $M_{jt,i}(s)$ be $\lambda_{jt,i}(s)$. Then, the likelihood for the observed claims process in (5.9) becomes:

$$\begin{aligned} L &= \prod_{j=1}^{J}\prod_{t=1}^{T_j}\left(\prod_{i=1}^{n_{jt}}\rho_{jt}(v_{jt,i})F_{U|v}(\tau - v_{jt,i})\exp\left(-\int_{t-1}^{t} w_{jt}(v)\rho_{jt}(v)F_{U|v}(\tau - v)dv\right)\right) \\ &\times \prod_{j=1}^{J}\prod_{t=1}^{T_j}\prod_{i=1}^{n_{jt}}\left(\prod_{k=1}^{m_{jt,i}}\lambda_{jt,i}(s_{jt,ik})\exp\left(-\int_{0}^{\tau_{jt,i}}\lambda_{jt,i}(s)ds\right)\right) \\ &\times \prod_{j=1}^{J}\prod_{t=1}^{T_j}\prod_{i=1}^{n_{jt}}\prod_{k=1}^{m_{jt,i}} f_P(p_{jt,ik}), \end{aligned} \tag{5.10}$$

where $\tau \in [T_j - 1, T_j]$, n_{jt} denote the number of reported claims that occur in the $\{jt\}$ observation period, and $m_{jt,i}$ is the number of transactions for claim i reported in the $\{jt\}$ observation period. With regards to the ratemaking application, the number of transactions to settlement is of interest. But for RBNS claims, the number of transactions is censored at the ratemaking date τ. Therefore, I denote, $\delta_{jt,i} = I(S_{jt,i} \leq \tau_{jt,i})$ to indicate whether the claim has been closed by the valuation time. Note that, $S_{jt,i}$ is the total waiting time from reporting to settlement of claim i reported in the $\{jt\}$ observation period. Thus, $\delta_{jt,i} = 1$ for closed claims, and $\delta_{jt,i} = 0$ for RBNS claims at τ. Additionally, given that $S_{jt,i} > \tau_{jt,i}$ for RBNS claims, it means that $M_{jt,i}(S_{jt,i}) \geq M_{jt,i}(\tau_{jt,i}) = m_{jt,i}$.

The likelihood in (5.10) for reported claims can be broken down into three building blocks: the number of claims per policy in a policy year, the conditional number of payment transactions for a claim, and the conditional payment sizes for each transaction. The likelihood is decomposed into independent blocks, which can be maximized in isolation. But, the MPP is a continuous-time model, and the data on the claims occurrence and transaction occurrence recorded and available for statistical inference are discrete. Thus,

I assume a piece-wise constant specification for the intensity functions that allow the use of the recorded number of claims and the number of transactions per claim for estimation. Each block is discussed below.

Poisson process for claim frequency N_{jt}: The first line in the likelihood in (5.10) relates to the occurrence of reported claims. A multiplicative form of the intensity function is assumed where $\rho_{jt}(v) = \rho_0(v;\alpha)\exp(x'_{jt}\beta)$. Here, x_{jt} are the rating variables described in Table 5.2, and $\{\alpha, \beta\}$ are parameters to be estimated. To estimate $\rho_{jt}(v)$, I assume the claim occurrence follows a Poisson process with a non-homogeneous piece-wise constant intensity ρ_{jt}, such that the baseline rate function is given by:

$$\rho_0(v;\alpha) = \alpha_t \quad a_{t-1} < v \leq a_t. \tag{5.11}$$

Here, $t = 1, \dots T$, and T is the most recent policy effective year. $\alpha = (\alpha_1, \dots, \alpha_T)$ are parameters of the baseline rate function, and $a_0 < a_1 < \cdots, a_T$ are the cut-points of the intervals for the baseline function where $a_0 = 0$ and $a_T = T$. Then, the occurrence of reported claims follows an independent Poisson process with intensity function $\rho_{jt}F_{U|v}(\tau - v)$. The corresponding likelihood for the occurrence of reported claims is given by:

$$L = \prod_{j=1}^{J}\prod_{t=1}^{T_j}\left(\rho_{jt}^{n_{jt}} \times \prod_{i=1}^{n_{jt}} F_{U|v}(\tau - v_{jt,i})\exp\left(-e_{jt}\rho_{jt}\int_{t-1}^{t} F_{U|v}(\tau - v)dv\right)\right), \tag{5.12}$$

where $\rho_{jt} = \alpha_t \exp(x'_{jt}\beta)$, and $e_{jt} = \int_0^\tau w_{jt}(u)du$ is the exposure time in $(a_{k-1}, a_k]$ for policyholder j. To optimize ρ_{jt}, the likelihood in (5.12) can be expressed without the reporting delay distribution function as a product of Poisson likelihoods shown as:

$$L = \prod_{j=1}^{J}\left(\prod_{t=1}^{T_j}(\alpha_t \exp(x'_{jt}\beta))^{n_{jt}}\exp\left(-e_{jt}\alpha_t \exp(x'_{jt}\beta)\right)\right). \tag{5.13}$$

Then, the likelihood in (5.13) can be maximized using a Poisson regression with a log

link where $\psi_{jt} = e_{jt} \exp(\log \alpha_t + x'_{jt}\beta)$ is the mean of the response N_{jt}. Here, $\ln(e_{jt})$ is specified as an offset variable to account for the exposure as at the ratemaking date.

Poisson process for transaction frequency $M_{jt,i}$: The second line in the likelihood in (5.10) relates to transaction occurrence conditional on having at least a claim. The transaction counting process $M_{jt,i}(s)$ is also Poisson with intensity measure $\lambda_{jt,i}(s) = \lambda_0(s;b)\exp(x'_{jt}\pi)$. Again, a piece-wise constant intensity $\lambda_{jt,i}$ is assumed such that the baseline rate function is given by:

$$\lambda_0(s;b) = b_t \quad a_{t-1} < s \leq a_t, \tag{5.14}$$

where, $b = (b_1, \ldots, b_T)$ are parameters of the baseline rate function. Then following a similar approach to that in (5.13), the likelihood for the transaction occurrence can be specified as a product of Poisson likelihoods shown as:

$$L = \prod_{j=1}^{J} \prod_{t=1}^{T_j} \left(\prod_{i=1}^{n_{jt}} (b_t \exp(x'_{jt}\pi))^{m_{jt,i}} \exp\left(-b_t \exp(x'_{jt}\pi)\right) \right). \tag{5.15}$$

For RBNS claims, $M_{jt,i}(S_{jt,i}) \geq m_{jt,i}$, then the likelihood in (5.15) can be maximized using censored Poisson regression with a log link where $\lambda_{jt,i} = \exp(\log b_t + x'_{jt}\pi)$ is the mean of the response $M_{jt,i}$. The likelihood for the censored Poisson is given by:

$$\Pr(m_{jt,i}, \delta_{jt,i}) = [f(m_{jt,i})]^{\delta_{jt,i}} \left[1 - \sum_{k=0}^{m_{jt,i}-1} f(k) \right]^{1-\delta_{jt,i}}, \tag{5.16}$$

where $f(\cdot)$ is a Poisson density function and $\delta_{jt,i} = 1$ if the claims is closed or $\delta_{jt,i} = 0$ if open.

Transaction severity $P_{jt,ik}$: The conditional severity block describes the claim payment size per transaction. I assume $P_{jt,ik}$ are i.i.d., and I specify a gamma regression with logarithmic link function and conditional mean $\mu_{jt,ik} = \exp(x'_{jt}\phi)$. Though I opted to

use the gamma GLM because it is frequently used in insurance pricing to model payment sizes (Henckaerts et al., 2018), other distributions can be used based on the data. In addition, following Antonio and Plat (2014), different models can be built for the first transaction payments and the later transaction payments.

5.3.2.3 IBNR Claims Modeling

The IBNR modeling block is considered as an additional building block of the MPP process. To account for the expected IBNR claims cost, I rely on a unique feature of the MPP model and define an IBNR factor:

$$\text{IBNR factor} = \frac{E[N(\tau)]}{E[N^{rep}(\tau)]}, \tag{5.17}$$

where $N(\tau)$ is the number of claims that occurred in an observation period $[0, \tau]$ and $N^{rep}(\tau)$ is the number of claims that occurred and reported by the valuation date τ. Then, under the Poisson process assumption, (5.17) becomes:

$$\begin{aligned} \text{IBNR factor} &= \frac{\int_0^\tau \rho(v)dv}{\int_0^\tau \rho(v)F_{U|v}(\tau-v)dv} \\ &= \frac{\int_0^\tau \rho(v)F_{U|v}(\tau-v)dv + \int_0^\tau \rho(v)(1-F_{U|v}(\tau-v))dv}{\int_0^\tau \rho(v)F_{U|v}(\tau-v)dv} \\ &= 1 + \frac{\int_0^\tau (1-F_{U|v}(\tau-v))dv}{\int_0^\tau F_{U|v}(\tau-v)dv}. \end{aligned} \tag{5.18}$$

Further, with the piece-wise constant intensity ρ_{jt} for the $\{jt\}$ observation period used in estimating the parameters in (5.13), the IBNR factor to account for all exposure for policyholder j is given by:

$$\text{IBNR factor}_j = \prod_{t=1}^{T_j} \left(1 + \frac{\int_{t-1}^t (1 - F_{U|v}(\tau - v))dv}{\int_{t-1}^t F_{U|v}(\tau - v)dv}\right). \tag{5.19}$$

An estimate of the IBNR factor is obtained by fitting the reporting delay distribution in (5.4).

5.3.2.4 How to Use the MPP for Ratemking

A rating formula based on the MPP ratemaking framework will be achieved by the product of the IBNR factor and the exponentiated estimates from the claim frequency model, transaction frequency model, and the severity model. The following rating formula calculates the predicted claims score:

$$\begin{aligned}\text{Premium} &= e_j \exp(\log \hat{\alpha}_T + x_j'\hat{\beta}) \times \exp(\log \hat{b}_T + x_j'\hat{\pi}) \times \exp(x_j'\hat{\phi}) \times \text{IBNR factor}_j \\ &= \text{Exposure} \times \text{Expected number of claims} \\ &\times \text{Expected number of transaction per claim} \\ &\times \text{Expected payment per transaction} \times \text{IBNR adjustment.}\end{aligned} \tag{5.20}$$

Where e_j is the exposure variable and x_j are rating factors for the new contact. $\{\hat{\alpha}_T, \hat{b}_T\}$ are the fitted baseline parameters from the most recent policy year for the claim and transaction frequency models. Also, $\{\hat{\beta}, \hat{\pi}, \hat{\phi}\}$ are the fitted parameters for rating variables from the claim frequency model, transaction frequency model, and the severity model building blocks. The loss reserves are automatically accounted for by using the information on open claims and adjusting for the IBNR claims. In this rating algorithm, the cumulative IBNR factor for new customers is one since there was no exposure in past years.

5.4 Estimation Results

In this section, I present the estimation results from the four building blocks in the MPP framework fitted using maximum likelihood. The training data contains observations from effective years 2006-2009, but I also show the parameter estimates using only data from the recent effective year 2009.

5.4.1 Claim Frequency Model

Table 5.4 presents the estimation results for the claim frequency Poisson model. The exposure variable is used as an offset, and the rating variables used are described in Table 5.2. The baseline parameters and the rating variables with parameter estimates that are not significant are not shown. When using all the training data, as expected, the coefficient for LnPolicyDed is negative, meaning higher deductible is associated with lower claim frequency, but the coefficient switches to positive and significant when using only observations from the effective year 2009. Also, the coefficients for the LnPolicyCov is positive and significant in both results. Compared to the reference category "Village," all entity types experience lower claims frequency except "Town" in both results. In addition, based on the Region rating factor, there are significant differences in claim frequency driven by the geographical location.

Table 5.4: Poisson claim frequency model parameter estimates.

	Effective Year 2009		Effective Years 2006-2009	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-5.090	0.272	-2.599	0.129
LnPolicyDed	0.214	0.035	-0.164	0.014
LnPolicyCov	1.148	0.049	1.110	0.021
TypeCity	-0.681	0.195	-0.535	0.076
TypeCounty	-0.892	0.206	-0.456	0.084
TypeSchool	-1.450	0.197	-1.005	0.076
TypeTown	0.050	0.474	0.415	0.149
TypeMisc	-2.349	0.364	-1.801	0.150
RegionNorthern	-0.280	0.180	0.151	0.084
RegionSoutheastern	-1.335	0.126	0.263	0.055
RegionSouthern	0.142	0.107	0.717	0.055
RegionWestern	-0.386	0.142	0.230	0.065
AlarmCredit(0%)	-	-	-0.076	0.063
AlarmCredit(5%)	-	-	-0.382	0.169
AlarmCredit(10%)	-	-	-0.306	0.132
AlarmCredit(15%)	-	-	0.008	0.046
-2 Log L	2,186		10,000	

5.4.2 Transaction Frequency Model

From Table 7.7 in Appendix 7.5, the transaction frequency did not vary much across the categorical variables, and that was confirmed in the censored Poisson transaction frequency model results in Table 5.5 as most of the parameter estimates were not statistically significant. Just like the claim frequency model, the LnPolicyDed is negative and significant when all the training data was used in the model building.

Table 5.5: Censored Poisson transaction frequency model parameter estimates.

	Effective Year 2009		Effective Years 2006-2009	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	0.097	0.042	0.290	0.045
LnPolicyDed	-	-	-0.024	0.009
-2 Log L	1,040		6,971	

5.4.3 Payment Severity Model

The estimation results for the Gamma severity model using a logarithmic link function is given in Table 5.6. The dependent variable is the observed transaction payments $P_{jt,ik}$ and the results show a significant difference in claim transaction payments based on geographical location. The results from all three building blocks show that some rating variables have a positive parameter estimate in one building block, but a negative parameter estimate in another. For example, using all the training dataset, the parameter estimate for LnPolicyDed is negative in the claim frequency model (-0.164), negative in the transaction frequency model (-0.024), and positive in the payment model (0.240). In this case, the overall effect can be interpreted as positive.

5.4.4 IBNR factor

The reporting delay is a key driver of IBNR claims. Figure 5.2 shows the distribution of the reporting delays in months overlayed with the fitted mixed distribution with a probability mass for a reporting delay of zero and a Weibull distribution for reporting

Table 5.6: Gamma severity model for average transaction payment.

	Effective Year 2009		Effective Years 2006-2009	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	11.662	0.435	7.822	0.138
LnPolicyDed	-0.426	0.045	0.290	0.020
LnPolicyCov	-	-	-0.208	0.021
RegionNorthern	0.966	0.317	0.759	0.119
RegionSoutheastern	1.920	0.187	0.332	0.080
RegionSouthern	0.583	0.194	0.156	0.080
RegionWestern	0.583	0.251	0.041	0.093
-2 Log L	9,826		71,543	

delays above zero. From the plot, the fitted mixed distribution seems to fit the observed reporting delay data reasonably well. Table 5.7 provides the IBNR factors based on (5.19) for each policy year without incorporating any covariates in the reporting delay distribution. From the results, using all the training data, the IBNR factor for the effective year 2009 is 1.210, and it means that $(1.210 - 1) \times 100 = 21\%$ of the reported claims are expected to be IBNR. The IBNR factors may vary depending on the policyholder characteristics; therefore, I expand the Weibull distribution for reporting delay to include the entity types. Table 5.8 shows the parameter estimates of the fitted Weibull model in (5.5).

Table 5.7: Estimates for IBNR factors without covariates.

	Effective Year 2009 IBNR factor	Effective Years 2006-2009 IBNR factor
2009	1.227	1.210
2008	-	1.006
2007	-	1.001
2006	-	1.000

Table 5.8: Weibull model parameter estimates for reporting delay.

	Estimate	Std.Error
(Intercept)	3.428	0.077
TypeCity	0.620	0.089
TypeCounty	0.979	0.089
TypeSchool	0.314	0.089
TypeTown	0.314	0.120
TypeMisc	0.908	0.198
$\log(\kappa)$	-0.285	0.014

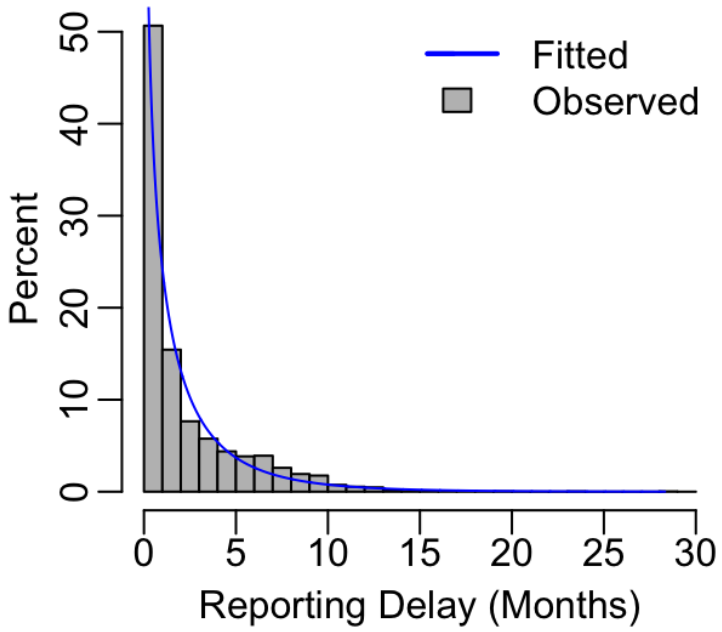

Figure 5.2: Observed reporting delay distribution overlayed with a fitted mixture of probability mass and Weibull Distributions (using all observations in the training data).

5.5 Out-of-Sample Performance

This section provides the claim score prediction based on the marked Poisson model fitted in Section 5.4. The predictions were generated based on the 2010 out-of-sample rating variables, and I compare the predictions to the 2010 out-of-sample claims. In addition, I compare the out-of-sample results to that of a frequency-severity model that uses only closed claims in the estimation of parameters named FreqSevClosed and another that uses all reported claims in the model building, named FreqSevAll. For the open claims in the FreqSevAll model, I use the incurred payment (amount paid plus loss reserve) as an estimate for the ultimate payment amount. In both frequency-severity models, the Poisson model is used to model the claim frequency and a gamma GLM with a logarithmic link is used to model the average payments using the number of claims as weights (Frees, 2014).

The Gini index measure developed in Frees et al. (2011) is employed to aid in the comparison of claim score predictions between the different models, and the out-of-sample claims. The Gini index is a measure of profit and defined as twice the average covariance between the predicted outcome and the rank of the predictor. Therefore, insurers that adopt a rating structure with a larger Gini index are more likely to enjoy a profitable portfolio. I show that the MPP framework helps align premium with the underlying risk better than the frequency-severity approach, consequently leading to a more profitable portfolio.

Table 5.9 presents the Gini index results. When the ratemaking model was built using data from the effective year 2009, the results show a smaller Gini index from the FreqSevClosed model as compared to the MPP framework. The results mean that the MPP framework promotes equity in pricing because it uses information on all reported claims and accounts for IBNR claims. The FreqSevAll model, which uses information on all reported claims, also performs better than the FreqSevClosed, highlighting the impact of information from open claims. Table 5.10 shows the differences in the Gini index for the models. It can be seen that the difference between the Gini indices between

the MPP framework and the FreqSevClosed model is small when the data from effective years 2006-2009 is used in the model building, which is because the proportion of closed claims increased.

Table 5.9: Gini indices of predictive claim scores.

	Effective Year 2009 Gini Index	Effective Years 2006-2009 Gini Index
FreqSevClosed	30.86%	68.86%
FreqSevAll	56.81%	68.31%
MPP	61.73%	67.39%

Table 5.10: Difference in Gini indices among scores.

Effective Year 2009		
	FreqSevAll	MPP
FreqSevClosed	25.95%	30.87%
FreqSevAll		4.92%
Effective Years 2006-2009		
	FreqSevAll	MPP
FreqSevClosed	-0.55%	-1.32%
FreqSevAll		-0.92%

Table 5.11 presents the Spearman correlations among predicted claim scores, and the out-of-sample claims. The MPP framework produced the highest correlation of 48.39% and 43.82% with the out-of-sample claims when models are based on data from the effective year 2009, and all the training dataset, respectively. This means that the MPP framework performs better than the frequency-severity models.

Table 5.11: Spearman correlations among scores and out of sample claims.

Effective Year 2009			
	FreqSevClosed	FreqSevAll	MPP
FreqSevAll	93.53%		
MPP	87.74%	86.42%	
Claims	38.49%	35.74%	48.39 %
Effective Years 2006-2009			
	FreqSevClosed	FreqSevAll	MPP
FreqSevAll	99.43%		
MPP	98.30%	98.01%	
Claims	42.87%	40.73%	43.82 %

Figure 5.3 shows the scatter plot between the out-of-sample claims and the predicted claim scores using logarithmic scaling. Each point on the plot represents a policyholder.

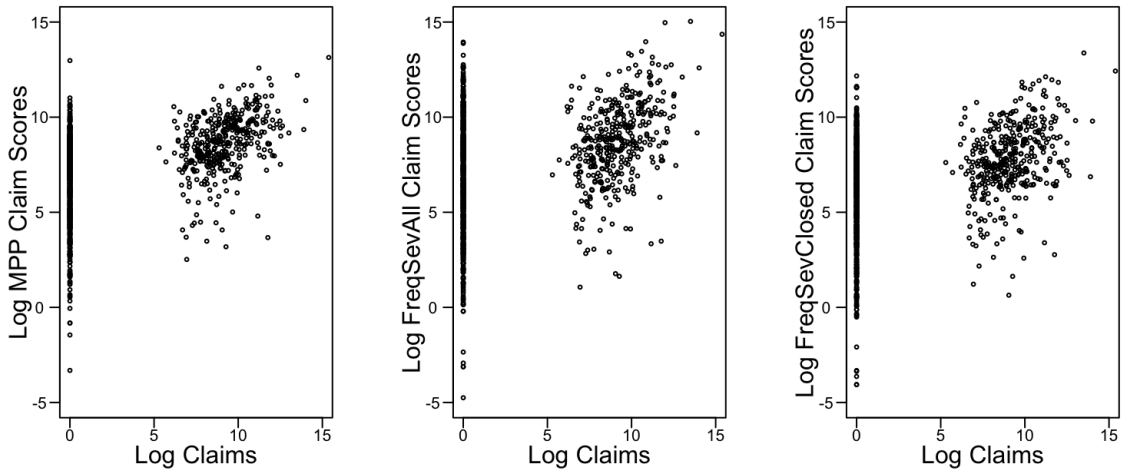

Figure 5.3: Comparison of claim scores (using data from the effective year 2009) to out-of-sample claims for 2010.

Because they are less spread out, the plot between the claims scores from the MPP model and the out-of-sample claims suggests higher correlations than the frequency-severity models.

Table 5.12 provides a robustness check for the results in Table 5.9. Here I use data from effective years 2007-2010 as the training dataset and use observations from the effective year 2011 as the hold-out-sample. Again, the Gini index using the MPP model is higher than using only the information on closed claims, which emphasizes the point that the MPP model promotes equity in rates and hence will lead to a more profitable portfolio.

Table 5.12: Gini indices of predictive claim scores for robustness check.

	Effective Year 2010 Gini Index	Effective Years 2007-2010 Gini Index
FreqSevClosed	58.77%	61.30%
FreqSevAll	56.88%	60.43%
MPP	66.67%	66.10%

5.6 Conclusion

Through the ratemaking process, insurance rates are set to cover the total future expected cost, which includes liabilities from both RBNS and IBNR claims. Actuaries develop rates by employing multivariate risk classification techniques based on information from the policy and the claim history to promote better alignment of premiums with claims experience. But the observation from the literature is that the data used in the multivariate analysis is usually based on closed claims, where the ultimate amount paid for all claims is known, leaving out open claims. Ignoring the information from open claims could lead to inaccurate rates because the ratemaking data lacks the current information that may capture shifts in the insurer's book risk profile.

This chapter employs the marked Poisson process (MPP) framework for ratemaking purposes by modeling four hierarchical building blocks. Three of the building blocks drive the expected cost based on reported claims by modeling the number of claims per policy in a policy year, the conditional number of payment transactions for a claim, and the conditional payment sizes for each transaction. Each block is modeled with the appropriate GLMs. The fourth building block account for the IBNR claims by deriving an IBNR factor based on the reported delay distribution. The results using data from a property insurance provider shows that the proposed framework promotes equity in the ratemaking algorithm.

Chapter 6

Summary and Concluding Remarks

This dissertation concerns using the individual-level claims prediction for insurance loss reserving and ratemaking. For loss reserving, complex cases can be both more expensive in terms of claims and take longer to settle, suggesting that the payment process is correlated with the settlement process for individual claims. In this case, knowledge of paid losses may help predict settlement time, which in turn feeds back into the prediction of unpaid losses. Further, when the settlement time and claim size are correlated, the historical claims that actuaries use for model building will not be representative of future payments, because large claims with longer settlement times will not be observed due to censoring by the valuation date, a type of selection bias.

I introduced a joint model framework to the individual-level loss reserving literature to accommodate such correlation. The joint model consists of a longitudinal submodel for the cumulative payment process and a survival submodel for the settlement process, and the correlation between the two components is induced via a shared parameter model. Macro-level reserving models like the chain-ladder technique are easy to implement, but they come with a risk of inaccurate predictions mainly because of their limited ability to handle claims heterogeneity and environmental or economic changes. However, the joint model incorporates both observed and unobserved heterogeneity into the two sub-processes, which is desired when one is interested in the prediction at the individual

claim level. Similarly, the joint model framework offers an improvement over the existing micro-level reserving models by explicitly accounting for the payment-settlement association, hence addressing the issue of selection bias.

To better understand the strength and limitations of the joint model for reserving applications, I present a detailed analysis using both simulated data and empirical data from a property insurance provider with the focus on RBNS claims. For the simulation study, I demonstrated that ignoring the payment-settlement association could lead to significant errors in reserving prediction. Further, I showed that the joint model could easily accommodate environmental changes such as a change in underwriting criteria, business mix, and claim processing, among others. However, the industry benchmark chain-ladder method without adjusting for the environmental changes produced a substantial error in reserving prediction. Moreover, I find that the advantages of the joint model are more pronounced for long-tail lines of business. In the empirical study, I provide a detailed analysis, which will make it easy for actuarial analysts to replicate the work. The prediction results from the joint model are compared to existing reserving models, and the results show that accounting for the payment-settlement association leads to better prediction accuracy. Because of the predictive nature of loss reserving, this dissertation enriches the existing statistical literature on joint models that have primarily focused on the estimation aspect of inference.

Historically, one main argument against micro-level models like the joint model over the years is that they are more difficult to implement in practice. However, with the growth in computing power in this era of big data analytics, insurance companies will enjoy a lesser burden with regards to implementation. Particularly, large companies with sophisticated personnel who are comfortable handling complex machine learning/AI type algorithms can implement and take advantage of micro-level models.

This dissertation concludes by employing the marked Poisson process (MPP) framework, which has primarily been used for micro-level reserving, to improve on insurance pricing. The MPP framework specified ensures that the multivariate risk analysis is done using

the information on claims that have been closed by the ratemaking date, payments on claims not yet closed, and reporting times to assess those claims that are IBNR. This work will be the first to provide a formal approach to incorporate information on open claims that have the ability to reflect shifts in the distribution of the expected claim payments. From an empirical study using data from a property insurance provider, I find that by allowing for current information, the proposed framework promotes equity in pricing and leads to a more profitable portfolio.

With the insurance industry experiencing a rapid pace of product innovation and intense competition, traditional ratemaking approaches that only employ closed claims from older policy years are inadequate. Further, the world is changing, and actuaries want to use the most recent information. In these pandemic times, change is everywhere. Recent experience means employing what information they can from open claims. By using the information on all reported claims, the proposed approach will also provide actuaries and regulators a more disciplined method of ascertaining promptly if rate increases are necessary.

Chapter 7

Appendix

7.1 Appendix to Chapter 1

7.1.1 Claims by Region

The 72 counties in Wisconsin are grouped into five regions using classifications from the Wisconsin Department of Health Services. Figure 7.1 shows claims by the region. It is seen that there are some variabilities in the number of claims and region.

7.1.2 Summary of Type of Payment Transactions

Table 7.1 summarizes the different payment transaction types in the training data.

Table 7.1: Type of payment transactions.

Transaction Type	Count
Payment Partial	323
Payment To Close	3,167
Payment After Close	4,01
Total	3,891

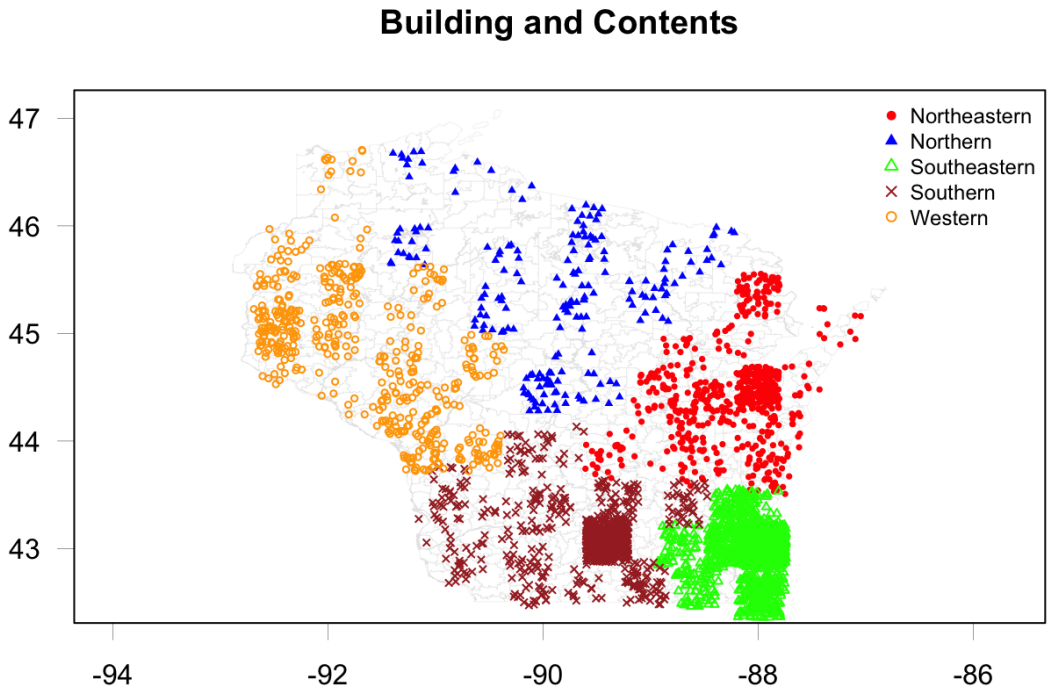

Figure 7.1: Claims by Region.

7.1.3 Loss Triangle from LGPIF Data for RBNS prediction

Table 7.2 summarizes the cumulative amounts paid arising out of building and contents coverage from the LGPIF data that occurred and were reported between January 1, 2006, and December 31, 2009, organized by reporting quarters and observation quarters. Then projections made from the developments factors give us RBNS reserve estimates. Table 7.3 provides the loss triangle without unusual claims.

Table 7.2: Observed historical cumulative claims $C_{i,j}$ organized by reporting quarters and observation quarters for RBNS prediction.

Reporting Quarter	Observation Quarter															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2006 Q1	1,264,447	1,682,188	1,682,188	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561
2006 Q2	2,009,311	3,370,260	4,761,115	5,174,541	5,873,088	6,559,705	6,818,143	6,818,143	7,003,825	7,003,825	7,003,825	7,003,825	7,003,825	7,003,825	7,003,825	
2006 Q3	1,402,769	2,406,855	2,482,578	2,523,994	2,615,403	2,670,567	2,670,567	2,670,567	2,670,567	2,670,567	2,670,567	2,670,567	2,670,567	2,670,567		
2006 Q4	806,480	1,207,453	1,239,388	1,572,304	4,227,033	4,227,033	4,227,033	4,481,423	4,481,423	4,481,423	4,481,423	4,481,423	4,481,423			
2007 Q1	1,135,006	1,788,184	1,917,445	2,123,798	2,590,466	2,654,564	2,654,564	2,720,276	2,720,276	2,720,276	2,720,276	2,720,276				
2007 Q2	705,146	1,350,025	1,841,137	2,010,819	6,004,549	6,232,118	6,619,945	6,696,295	8,035,846	8,035,846	8,035,846					
2007 Q3	1,100,841	1,854,040	2,193,211	2,234,338	2,612,670	2,637,070	2,637,070	2,637,070	2,671,639	2,744,121						
2007 Q4	1,893,020	2,951,579	3,380,631	3,725,657	3,756,556	3,758,990	3,765,436	3,953,195	3,976,862							
2008 Q1	1,488,889	2,944,819	3,234,739	3,675,551	3,696,906	3,812,083	4,787,062	5,041,970								
2008 Q2	1,516,620	3,537,807	4,642,772	5,173,249	5,209,948	5,249,361	5,249,361									
2008 Q3	1,375,480	2,860,584	3,612,216	3,762,930	3,792,355	3,796,699										
2008 Q4	1,046,145	1,537,226	1,805,164	1,956,216	1,956,216											
2009 Q1	1,277,018	1,779,555	1,971,026	2,058,902												
2009 Q2	816,927	1,625,055	1,810,738													
2009 Q3	1,396,415	1,816,822														
2009 Q4	450,633															

Table 7.3: Observed historical cumulative claims $C_{i,j}$ organized by reporting quarters and observation quarters for RBNS prediction (without unusual claims).

Reporting Quarter	Observation Quarter															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2006 Q1	1,264,447	1,682,188	1,682,188	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561	1,722,561
2006 Q2	2,009,311	3,370,260	4,761,115	5,174,541	5,873,088	6,559,705	6,818,143	6,818,143	7,003,825	7,003,825	7,003,825	7,003,825	7,003,825	7,003,825	7,003,825	
2006 Q3	1,402,769	2,406,855	2,482,578	2,523,994	2,615,402	2,670,566	2,670,566	2,670,566	2,670,566	2,670,566	2,670,566	2,670,566	2,670,566	2,670,566		
2006 Q4	806,480	1,207,453	1,239,388	1,572,304	4,227,033	4,227,033	4,227,033	4,481,423	4,481,423	4,481,423	4,481,423	4,481,423	4,481,423			
2007 Q1	1,135,006	1,788,184	1,917,445	2,123,798	2,590,466	2,654,564	2,654,564	2,720,276	2,720,276	2,720,276	2,720,276	2,720,276				
2007 Q2	601,388	1,246,267	1,570,116	1,739,799	2,107,930	2,335,498	2,335,498	2,411,848	2,637,794	2,637,794	2,637,794					
2007 Q3	1,100,841	1,854,040	2,193,211	2,234,338	2,612,670	2,637,070	2,637,070	2,637,070	2,671,639	2,744,121						
2007 Q4	1,893,020	2,951,579	3,380,631	3,725,657	3,756,556	3,758,989	3,765,435	3,953,195	3,976,862							
2008 Q1	1,488,889	2,363,739	2,653,659	3,094,471	3,115,826	3,231,003	3,239,227	3,239,227								
2008 Q2	1,516,620	3,537,807	4,642,772	5,173,249	5,209,948	5,249,361	5,249,361									
2008 Q3	1,375,480	2,860,584	3,612,216	3,762,930	3,792,355	3,796,699										
2008 Q4	1,046,145	1,537,226	1,805,164	1,956,216	1,956,216											
2009 Q1	1,277,018	1,779,555	1,971,026	2,058,902												
2009 Q2	816,927	1,625,055	1,810,738													
2009 Q3	1,396,415	1,816,822														
2009 Q4	450,633															

7.1.4 Loss Triangle from LGPIF Data for Total Liabilities Prediction

Table 7.4 summarizes the cumulative amounts paid arising out of building and contents coverage from the LGPIF data that occurred and were reported between January 1, 2006, and December 31, 2009, organized by accident quarters and development quarters. Then projections made from the development factors give us the total reserve estimates.

Table 7.4: Observed historical cumulative claims $C_{i,j}$ organized by accident quarters and development quarters for total liabilities prediction.

Accident Quarter	Development Quarter															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2006 Q1	785,132	1,596,392	1,876,498	2,016,610	2,042,924	2,070,222	2,076,592	2,076,592	2,076,592	2,076,592	2,082,121	2,082,121	2,082,121	2,082,121	2,082,121	2,082,121
2006 Q2	399,367	2,721,031	4,777,790	5,321,127	6,124,351	6,236,634	6,910,455	7,140,728	7,140,728	7,326,411	7,326,411	7,326,411	7,326,411	7,326,411	7,326,411	
2006 Q3	389,923	1,849,309	2,505,428	2,669,470	2,759,763	2,881,102	2,881,102	2,881,102	2,883,678	2,883,678	2,883,678	2,883,678	2,883,678	2,883,678		
2006 Q4	81,824	603,722	917,238	1,183,285	1,609,461	4,131,248	4,131,248	4,385,639	4,385,639	4,385,639	4,385,639	4,385,639	4,385,639			
2007 Q1	260,093	1,364,663	1,647,783	1,826,755	2,209,008	2,381,272	2,436,895	2,436,895	2,502,607	2,502,607	2,502,607	2,502,607				
2007 Q2	184,181	809,595	1,358,734	1,997,094	5,979,528	6,303,318	6,788,504	6,864,854	7,090,801	8,204,405	8,204,405					
2007 Q3	316,584	1,579,128	2,050,476	2,132,666	2,600,897	2,639,768	2,639,768	2,639,768	2,639,768	2,746,819						
2007 Q4	452,496	2,259,561	3,296,680	3,462,907	3,804,200	3,898,454	3,906,732	3,914,956	4,126,383							
2008 Q1	384,822	2,406,894	2,937,926	3,519,804	3,573,898	3,689,991	3,959,055	4,911,882								
2008 Q2	235,470	2,934,045	4,484,925	5,965,053	6,144,670	6,196,276	6,239,803									
2008 Q3	282,890	2,063,774	2,958,419	3,109,119	3,181,833	3,205,751										
2008 Q4	146,419	936,693	1,276,588	1,426,504	1,512,838											
2009 Q1	106,139	1,022,974	1,407,795	1,738,040												
2009 Q2	415,360	1,084,856	1,792,756													
2009 Q3	344,604	1,663,680														
2009 Q4	214,008															

7.2 Appendix to Chapter 2

In this section, alternative estimating strategies for the joint model are discussed. It is shown in Tables 3.2 and 3.3 that significant bias will be induced using these techniques.

7.2.1 Independent Estimation

The independent estimation ignores the payment-settlement association. Specifically, setting $\alpha = 0$ in the survival submodel, and the longitudinal and survival submodel are estimated separately. The hazard function of time-to-settlement outcome T_i^* of a claim is modeled using a proportional hazards model specified as:

$$h_i(t) = h_0(t) \exp\{\gamma' \mathbf{w}_{it}\}, \tag{7.1}$$

where $h_0(t)$ is the baseline hazard function, and w_{it} is a vector of covariates with a corresponding vector of regression coefficients γ. Under this independent estimation, the longitudinal process is modeled using the GLMM specification in (2.4).

7.2.2 Two-Stage Estimation

The two-stage approach attempts to incorporate the payment-settlement association. The first stage estimates the longitudinal submodel, and the second stage estimates the survival submodel holding parameter estimates from the first stage fixed. Then the hazard function of time-to-settlement outcome T_i^* of a claim is modeled using a proportional hazards model specified as:

$$h_i(t) = h_0(t) \exp\{\gamma' \mathbf{w}_{it} + \alpha \eta_{it}\}, \tag{7.2}$$

The effect longitudinal cumulative payments on the risk of settlement is given by the parameter α. Just like the independent estimation, the longitudinal process is modeled using the GLMM specification in (2.4).

7.3 Appendix to Chapter 3

7.3.1 Sample R Code for Joint Model Simulation, Estimation and Prediction

This code is based on work by Sweeting and Thompson (2011).

```
# Simulate a joint model assuming a Linear Mixed-effects
# for the longitudinal model and a Cox proportional
# hazard model with an exponential baseline function
# for the survival submodel.

library(JM)
library(MASS)
library(Hmisc)
```

```r
library(foreign)
library(flexsurv)
library(devtools)
#install_github("jwdink/tidysurv")
library(tidysurv)
library(data.table)
library(stringi)

#Parameters
n<-1000          # Number of claims
k<-11            # Number of payment times for each claim
Sigma<-1.5       # Random effect sd
beta<-c(1,0.5) # Time intercept and slope payment trend
beta3<-0.4       # effect of X1 ( a binary variable) on long. submodel
beta4<-0.3       # effect of X2 ( a continous variable) on long. submodel
sigma0<-1.5      # standard error for longitudinal model
shape<-1         # shape for weilbull ( exponential) - baseline hazard
rate<-0.4        # scale for weilbull - baseline hazerd
alpha<--0.25     # association parameter
gam1<-0.5        # effect of X1 on survival submodel
gam2<-0.3        # effect of X2 on survival submodel
seed=1231        # seed for reproducibility
max.time=10      # maximum time simulated

# Function for simulation
## Baseline survival function
sc<-function(times=NULL){
  exp(-rate * times^shape)}

## hazard ratio
```

```
hr<-function(times=NULL){
  mu<-function(t){ cbind(1,t)%*%t(t(beta)) + b[i]+
      beta3*data.surv$X1[i]+
      beta4*data.surv$X2[i]}
  exp(alpha*(mu(times))+gam1*(data.surv$X1[i])+gam2*(data.surv$X2[i]))
}

## Step 1:  Simulate longitudinal cumulative dataset
set.seed(seed)
data.long <- data.frame(id = rep(1:n, each = k),
                        time = rep(seq(0, max.time, length=k), n),
                        X1=rep(rbinom(n, 1, 0.5), each = k),
                        X2=rep(rnorm(n,1,0.5), each = k))
b <-rnorm(n, 0, Sigma)

data.long$AccTime<-trunc((data.long$id-1)/100)
data.long$CurTime<-data.long$AccTime+data.long$time
data.long$y <- rnorm(n*k,
                     mean =  cbind(1,data.long$time)%*%t(t(beta)) +
                       b[data.long$id]+
                       beta3*data.long$X1+
                       beta4*data.long$X2,
                     sd = sigma0)

##Step 2: Simulate time to settlement for each claim
data.surv<-data.frame(id=1:n,
                      random.cens=runif(n,0,10),
                      X1=data.long$X1[ !duplicated(data.long$id) ],
                      X2=data.long$X2[ !duplicated(data.long$id) ])
data.longsub<-subset(data.long,select = c(id,AccTime))
```

```
data.longsub<-data.longsub[!duplicated(data.longsub$id, fromLast = T), ]
data.surv<-merge(data.surv,data.longsub, by="id", all.x = T)
data.surv$true.time<-NA
  rint<-function(n) surv(n,what='int')
  rcon<-function(n) surv2(n,what='control')
  for(i in 1:n){
    surv<-Quantile2(sc,hr,mplot=1000,tmax=11,pr=FALSE)
    data.surv$true.time[i]<-rint(1)
  }
  data.surv$true.time<-as.numeric(data.surv$true.time)
data.survSub<-subset(data.surv,select = c(id,true.time))
data.long<-merge(data.long,data.survSub, by="id", all.x = T)
data.long$random.cens<-ifelse(data.long$AccTime==0,9,
                              ifelse(data.long$AccTime==1,8,
                              ifelse(data.long$AccTime==2,7,
                              ifelse(data.long$AccTime==3,6,
                              ifelse(data.long$AccTime==4,5,
                              ifelse(data.long$AccTime==5,4,
                              ifelse(data.long$AccTime==6,3,
                              ifelse(data.long$AccTime==7,2,
                              ifelse(data.long$AccTime==8,1,
                              ifelse(data.long$AccTime==9,0,NA)))))))))
data.long$random.cens<-data.long$random.cens+rep(runif(n,0,0.5),each= k)
data.surv<-subset(data.surv,select = -c(random.cens))
data.longsub<-subset(data.long,select = c(id,random.cens))
data.longsub<-data.longsub[!duplicated(data.longsub$id, fromLast = T), ]
data.surv<-merge(data.surv,data.longsub, by="id", all.x = T)
data.surv$event<-ifelse(data.surv$random.cens<data.surv$true.time,0,1)
data.surv$event.time<-pmin(data.surv$random.cens,data.surv$true.time)
data.survSub<-subset(data.surv,select = c(id,event.time))
```

```
data.long<-merge(data.long,data.survSub, by="id", all.x = T)

##Step 3: Remove longitudinal data with time of payments
## greater than min(settlement time,valuation time)
data.longU<-subset(data.long,data.long$time>data.long$event.time)
data.long<-subset(data.long,data.long$time<=data.long$event.time)

## Remove claims from survival data-set whose event time is zero
data.surv<-subset(data.surv,data.surv$event.time!=0)
data.long<-subset(data.long,data.long$id %in%  data.surv$id )

## Adding stopping times for interval data (for two-stage modeling)
data.long$stop<-c(data.long$time[-1],0)
data.long$stop[data.long$stop==0]<-data.long$event.time[data.long
                                                        $stop==0]

### Event indicator for longitudinal data (for two-stage modeling)
data.long$event<-0
data.long$event[unlist(by(data.long,factor(data.long$id),
                  function(i){i$time==max(i$time)}))]<-data.surv$event

# Estimation
## Independent model
### Longitudinal submodel: LMM with randam intercept
LMEfit<-lme(y~time + X1+X2,random=~1|id,data=data.long,method="ML")

### Parmetric Cox survival model with exponentail baseline
marginal.surv <-flexsurvreg(Surv(event.time, event) ~ X1+X2 ,
                            data = data.surv,dist = "exp")
```

```
## Two stage
### Stage 1 : LMM with randam intercept
LMEfit<-lme(y~time + X1+X2,random=~1|id,data=data.long,method="ML")
data.long$y_pred<-predict(LMEfit)

### Stage 2: Extended Cox model with exponentail baseline
marginal.survTS<- flexsurvreg(Surv(time, stop, event) ~ X1+X2 + y_pred,
                              data = data.long,dist = "exp")

## Joint model assuming Normal distribution for longitudinal dataset
## and an exponential baseline survival model (scale=1)
# Provide starting values for longitudinal submodel
LMEfit<-lme(y~time + X1+X2,random=~1|id,data=data.long,method="ML")

# Provide starting values for survival submodel
marginal.survJM<-survreg(Surv(event.time, event) ~ X1+X2,
                         data = data.surv, x = TRUE,scale=1)

marginal.survJM$y<-as.matrix(cbind(log(as.numeric(ifelse(
  stri_sub(marginal.survJM$y,-1)=="+",stri_sub(
    marginal.survJM$y,1,-2),stri_sub(marginal.survJM$y,1,-1))
)),as.numeric(ifelse(stri_sub(marginal.survJM$y,-1)
                      =="+",0,1))))

## Joint Model Estimation Using JM package in R
jointModel_fit<-jointModel(LMEfit,marginal.survJM,
                           timeVar="time",method="weibull-PH-GH",scale=1)
```

```
# Prediction
## Create validation dataset containing open claims for prediction
data.survT<-data.surv
data.survT$time<-ifelse(data.survT$true.time<1,0,
      ifelse(data.survT$true.time>=1 & data.survT$true.time<2,1,
      ifelse(data.survT$true.time>=2 & data.survT$true.time<3,2,
      ifelse(data.survT$true.time>=3 & data.survT$true.time<4,3,
      ifelse(data.survT$true.time>=4 & data.survT$true.time<5,4,
      ifelse(data.survT$true.time>=5 & data.survT$true.time<6,5,
      ifelse(data.survT$true.time>=6 & data.survT$true.time<7,6,
      ifelse(data.survT$true.time>=7 & data.survT$true.time<8,7,
      ifelse(data.survT$true.time>=8 & data.survT$true.time<9,8,
      ifelse(data.survT$true.time>=9 & data.survT$true.time<10,9,9
             ))))))))))
data.survT<-subset(data.survT,event==0&AccTime!=0 )
data.longT<-subset(data.long,data.long$id %in%  data.survT$id )

data.survTUlt<-subset(data.survT,select = c(id,time))
colnames(data.survTUlt)<-c("id","timeUlt")
data.longU<-subset(data.longU,data.longU$id %in%  data.survT$id )
data.longU<-merge(data.longU,data.survTUlt, by="id", all.x = T)
data.longU<-subset(data.longU,data.longU$time==data.longU$timeUlt)
data.longU$yUlt<-data.longU$y
data.survT<-subset(data.survT,data.survT$id %in% data.longU$id )
data.longU<-subset(data.longU,select = c(id,yUlt))
data.survT<-merge(data.survT,data.longU, by="id", all.x = T)
data.longT<-subset(data.longT,data.longT$id %in% data.longU$id )
data.survTsub<-subset(data.survT, select = c(id,yUlt))
data.longT<-merge(data.longT,data.survTsub, by="id", all.x = T)
```

```
## Independent model
if(class(LMEfit)!="try-error" & class(marginal.surv)!="try-error"){
  simulTimeInd <- function(survmodel, data, SurvCen)
  {
    ND=data
    rate<- exp(survmodel$coefficients[1])
    gam1<- survmodel$coefficients[2]
    gam2<- survmodel$coefficients[3]
    shape=1
    SurvCen=SurvCen
    set.seed(5)
    v <- runif(n=10000)
    v<-v*SurvCen
    Tlat <- (- log(v) / (rate * exp(gam1*(ND$X1[1])
                          +gam2*(ND$X2[1]))))^(1 / shape)
    return(Tlat)
  }
  data.longT1<-data.longT[!duplicated(data.longT$id, fromLast = T), ]
  N<-dim(data.longT1)[1]
  TestRBNSPaidPredZ=data.frame(cbind(rep(0,N),rep(0,N),rep(0,N),
rep(0,N)))
  colnames(TestRBNSPaidPredZ)<-c("id","Pred","TrueTime","PredTime1")
  TestRBNSPaidPredZ$id<-data.longT1$id
  TestRBNSPaidPredZ$yUlt<-data.longT1$yUlt
  RanEff=data.frame(cbind(rep(0,dim(data.surv)[1]),
rep(0,dim(data.surv)[1])))
  colnames(RanEff)<-c("id","ranefInt")
  RanEff$id<-data.surv$id
  RanEff$ranefInt<-ranef(LMEfit)
```

```
uniq <- unique(unlist(TestRBNSPaidPredZ$id))
for (i in 1:length(uniq)){
  ND <- subset(data.longT1, id == uniq[i])
  NDcum<-ND[!duplicated(ND$id, fromLast = T), ]
  CumLast<-NDcum$y
  survTimes<-seq(ND$random.cens[1],11, length.out = 35)
  intervals <- data.frame(id=rep(ND$id, 35), time=survTimes)
  covs <- ND[ c("X1", "X2","true.time")]
  ND<-data.frame(covs, intervals, row.names = NULL)
  # survival probability at censoring/valuation time
  predSurvZProb<-predict(marginal.surv,
                   newdata=ND[1,], ND$time[1], type = "survival")
  SurvCen<-predSurvZProb
  # conditional survival distribution after censoring/valuation time
  TlatPred <- simulTimeInd(survmodel=marginal.surv,
                        data=ND,SurvCen=SurvCen)
  # Predicted settlement time
  Times<-mean(TlatPred,na.rm=T)

  Times<-ifelse(is.na(Times),9,Times)
  Times<-ifelse(Times>=9,9,Times)
  True.time<-ifelse(ND$true.time[1]>=9,9,ND$true.time[1])
  ND$time<-Times
  raneffi<-as.numeric(RanEff[which(RanEff[,1]==uniq[i]),2])
  raneff=raneffi
  # Ultimate payment prediction
  lfit <- predict(LMEfit, newdata = ND, level = 0)[1]
  mean<-lfit+raneff
  TestRBNSPaidPredZ[i,2]<-mean
```

```
    TestRBNSPaidPredZ[i,3]<- True.time
    TestRBNSPaidPredZ[i,4]<- Times
  }
  data.longTsub<-data.longT[!duplicated(data.longT$id, fromLast = T), ]
  data.longTsub<-subset(data.longTsub,select = c(id,y))
  colnames(data.longTsub)<-c("id","y_t")
  TestRBNSPaidPredZ<-merge(TestRBNSPaidPredZ,data.longTsub, by="id")
  TestRBNSPaidPredZ$Reserve<-TestRBNSPaidPredZ$Pred-
TestRBNSPaidPredZ$y_t
  TestRBNSPaidPredZ$ActReserve<-TestRBNSPaidPredZ$yUlt-
TestRBNSPaidPredZ$y_t
  TotalReserve<-sum(TestRBNSPaidPredZ$ActReserve)
  ErrorInd<-sum(TestRBNSPaidPredZ$Reserve)-
sum(TestRBNSPaidPredZ$ActReserve)
} else {
  data.longT1<-data.longT[!duplicated(data.longT$id, fromLast = T), ]
  N<-dim(data.longT1)[1]
  TestRBNSPaidPredZ=data.frame(cbind(rep(0,N),rep(0,N)))
  colnames(TestRBNSPaidPredZ)<-c("id","Pred")
  TestRBNSPaidPredZ$id<-data.longT1$id
  TestRBNSPaidPredZ$yUlt<-data.longT1$yUlt
  data.longTsub<-data.longT[!duplicated(data.longT$id, fromLast = T), ]
  data.longTsub<-subset(data.longTsub,select = c(id,y))
  colnames(data.longTsub)<-c("id","y_t")
  TestRBNSPaidPredZ<-merge(TestRBNSPaidPredZ,data.longTsub, by="id")
  TestRBNSPaidPredZ$ActReserve<-TestRBNSPaidPredZ$yUlt-
TestRBNSPaidPredZ$y_t
  TotalReserve<-sum(TestRBNSPaidPredZ$ActReserve)
  ErrorInd<-NA
}
```

```
## Two stage
PredSurvTS <- function(longmodel,survmodel, data, raneffi,ranefft)
{
  ND=data
  beta1<- fixed.effects(longmodel)[1]
  betat<-fixed.effects(longmodel)[2]
  sigma0<-longmodel$sigma
  shape<-1
  rate<- survmodel$res[1,1]
  alpha<-survmodel$res[4,1]
  beta3<-fixed.effects(longmodel)[3]
  beta4<-fixed.effects(longmodel)[4]
  gam1<- survmodel$res[2,1]
  gam2<- survmodel$res[3,1]
  b0<-raneffi
  bt<-ranefft
  time=ND$time
  first<- rate*(exp((gam1*(ND$X1[1])+gam2*(ND$X2[1]))
                    +alpha*(beta1+b0+beta3*ND$X1[1]+beta4*ND$X2[1])))
  second<-alpha*(betat+bt)
  third<- (exp(time*(alpha*(betat+bt))))-1
  CumHaz<- (first/second)*third
  Surv<-exp(-CumHaz)
  return(Surv)
}

simulTimeTS <- function
(longmodel,survmodel, data, raneffi,ranefft, SurvCen)
```

```
{
  ND=data
  beta1<- fixed.effects(longmodel)[1]
  betat<-fixed.effects(longmodel)[2]
  sigma0<-longmodel$sigma
  shape<-1
  rate<- survmodel$res[1,1]
  alpha<-survmodel$res[4,1]
  beta3<-fixed.effects(longmodel)[3]
  beta4<-fixed.effects(longmodel)[4]
  gam1<- survmodel$res[2,1]
  gam2<- survmodel$res[3,1]
  b0<-raneffi
  bt<-ranefft
  SurvCen=SurvCen
  set.seed(5)
  v <- runif(n=10000)
  v<-v*SurvCen
  first<- -log(v)*(alpha*(betat+bt))
  second<- rate*(exp((gam1*(ND$X1[1])+gam2*(ND$X2[1]))
                     +alpha*(beta1+b0+beta3*ND$X1[1]+beta4*ND$X2[1])))
  third<- alpha*(betat+bt)
  Tlat <- (log((first/ second)+1))/third
  return(Tlat)
}



if(class(LMEfit)!="try-error"){
  data.long$y.cen_pred<-predict(LMEfit)
  marginal.survTS<- try(flexsurvreg(Surv(time, stop, event) ~ X1+X2
```

```
                    + y.cen_pred, data = data.long,dist = "exp"))

  if(class(marginal.survTS)!="try-error"){
   data.longT1<-data.longT[!duplicated(data.longT$id, fromLast = T), ]
    N<-dim(data.longT1)[1]
    TestRBNSPaidPredZ=data.frame(cbind(rep(0,N),rep(0,N),rep(0,N),
rep(0,N)))
    colnames(TestRBNSPaidPredZ)<-c("id","Pred","TrueTime","PredTime1")
    TestRBNSPaidPredZ$id<-data.longT1$id
    TestRBNSPaidPredZ$yUlt<-data.longT1$yUlt
    RanEff=data.frame(cbind(rep(0,dim(data.surv)[1]),rep(0,
dim(data.surv)[1])))
    colnames(RanEff)<-c("id","ranefInt")
    RanEff$id<-data.surv$id
    RanEff$ranefInt<-ranef(LMEfit)

    uniq <- unique(unlist(TestRBNSPaidPredZ$id))
    for (i in 1:length(uniq)){
      ND <- subset(data.longT1, id == uniq[i])
      NDcum<-ND[!duplicated(ND$id, fromLast = T), ]
      CumLast<-NDcum$y
      lasttime=ND$random.cens
      survTimes<-seq(ND$random.cens[1],11, length.out = 35)
      intervals <- data.frame(id=rep(ND$id, 35), time=survTimes)
      covs <- ND[ c("X1", "X2","true.time")]
      ND<-data.frame(covs, intervals, row.names = NULL)
      ND$y.cen_pred<-predict(LMEfit, newdata = ND)
      raneffi<-as.numeric(RanEff[which(RanEff[,1]==uniq[i]),2])
      # survival probability at censoring/valuation time
      predSurvZProb<-PredSurvTS(longmodel=LMEfit,
```

```
                        survmodel=marginal.survTS, data=ND[1,],
                        raneffi= raneffi,ranefft= 0)
      SurvCen<-predSurvZProb
      # conditional survival distribution after censoring/valuation
      #  time
      TlatPred<-simulTimeTS(longmodel=LMEfit,survmodel=marginal.survTS,
                        data=ND[1,], raneffi= raneffi,
                        ranefft= 0, SurvCen=SurvCen)
      # Predicted settlement time
      Times<-mean(TlatPred,na.rm=T)
      Times<-ifelse(is.na(Times),9,Times)
      Times<-ifelse(Times>=9,9,Times)
      ND<-ND[!duplicated(ND$id, fromLast = T), ]
      True.time<-ifelse(ND$true.time[1]>=9,9,ND$true.time[1])
      ND$time<-Times
      raneffi<-as.numeric(RanEff[which(RanEff[,1]==uniq[i]),2])
      raneff=raneffi
      # Ultimate payment prediction
      lfit <- predict(LMEfit, newdata = ND, level = 0)[1]
      mean<-lfit+raneff
      TestRBNSPaidPredZ[i,2]<-mean
      TestRBNSPaidPredZ[i,3]<- True.time
      TestRBNSPaidPredZ[i,4]<- Times
    }
    data.longTsub<-data.longT[!duplicated(data.longT$id, fromLast = T),]
    data.longTsub<-subset(data.longTsub,select = c(id,y))
    colnames(data.longTsub)<-c("id","y_t")
    TestRBNSPaidPredZ<-merge(TestRBNSPaidPredZ,data.longTsub, by="id")
    TestRBNSPaidPredZ$Reserve<-TestRBNSPaidPredZ$Pred-
TestRBNSPaidPredZ$y_t
```

```
    TestRBNSPaidPredZ$ActReserve<-TestRBNSPaidPredZ$yUlt-
TestRBNSPaidPredZ$y_t
    Error2S<-sum(TestRBNSPaidPredZ$Reserve)-
sum(TestRBNSPaidPredZ$ActReserve)
} else {
    Error2S<-NA
  }

}else {
  Error2S<-NA
}



## Joint model
if(class(jointModel_fit)!="try-error"){
  H <- jointModel_fit$Hessian
  ev <- eigen(H, symmetric = TRUE, only.values = TRUE)$values
  if (!all(ev >= 1e-08 )|any(is.na(H) | !is.finite(H))){
    ErrorJM<-NA
  } else {
    PredSurvJM <- function(model, data, raneffi,ranefft)
    {
      ND=data
      beta1<- model$coefficients$betas[1]
      betat<-model$coefficients$betas[2]
      sigma0<-model$coefficients$sigma
      shape<-1
      rate<- exp(model$coefficients$gammas[1])
      alpha<-model$coefficients$alpha
      beta3<-model$coefficients$betas[3]
```

```
  beta4<-model$coefficients$betas[4]
  gam1<- model$coefficients$gammas[2]
  gam2<- model$coefficients$gammas[3]
  b0<-raneffi
  bt<-ranefft
  time=ND$random.cens[1]
  first<- rate*(exp((gam1*(ND$X1[1])+gam2*(ND$X2[1]))+
                alpha*(beta1+b0+beta3*ND$X1[1]+beta4*ND$X2[1])))
  second<-alpha*(betat+bt)
  third<- (exp(time*(alpha*(betat+bt))))-1
  CumHaz<- (first/second)*third
  Surv<-exp(-CumHaz)
  return(Surv)
}

simulTime <- function(model, data, raneffi,ranefft, SurvCen)
{
  ND=data
  beta1<- model$coefficients$betas[1]
  betat<-model$coefficients$betas[2]
  sigma0<-model$coefficients$sigma
  shape<-1
  rate<- exp(model$coefficients$gammas[1])
  alpha<-model$coefficients$alpha
  beta3<-model$coefficients$betas[3]
  beta4<-model$coefficients$betas[4]
  gam1<- model$coefficients$gammas[2]
  gam2<- model$coefficients$gammas[3]
  b0<-raneffi
  bt<-ranefft
```

```
      SurvCen=SurvCen
      set.seed(5)
      v <- runif(n=10000)
      v<-v*SurvCen
      first<- -log(v)*(alpha*(betat+bt))
      second<- rate*(exp((gam1*(ND$X1[1])+gam2*(ND$X2[1]))+
                alpha*(beta1+b0+beta3*ND$X1[1]+beta4*ND$X2[1])))
      third<- alpha*(betat+bt)
      Tlat <- (log((first/ second)+1))/third
      return(Tlat)
    }

    data.longT1<-data.longT[!duplicated(data.longT$id, fromLast = T), ]
    N<-dim(data.longT1)[1]
    TestRBNSPaidPredZ=data.frame(cbind(rep(0,N),rep(0,N),rep(0,N),
rep(0,N)))
    colnames(TestRBNSPaidPredZ)<-c("id","Pred","TrueTime","PredTime1")
    TestRBNSPaidPredZ$id<-data.longT1$id
    TestRBNSPaidPredZ$yUlt<-data.longT1$yUlt
    RanEff=data.frame(cbind(rep(0,dim(data.surv)[1]),rep(0,
dim(data.surv)[1])))
    colnames(RanEff)<-c("id","ranefInt")
    RanEff$id<-data.surv$id
    RanEff$ranefInt<-ranef(jointModel_fit)

    uniq <- unique(unlist(TestRBNSPaidPredZ$id))
    for (i in 1:length(uniq)){
      ND <- subset(data.longT, id == uniq[i])
      NDcum<-ND[!duplicated(ND$id, fromLast = T), ]
      CumLast<-NDcum$y
```

```
if(dim(ND)[1] == 1){
  ND$time<-0.1
}
if( ND$random.cens[1] == 0){
  ND$random.cens<-0.1
}
lasttime=ND$random.cens[1]
raneffi<-as.numeric(RanEff[which(RanEff[,1]==uniq[i]),2])
# survival probability at censoring/valuation time
SurvCen<-PredSurvJM(model=jointModel_fit, data=ND[1,],
                raneffi=raneffi,ranefft=0)
# conditional survival distribution after censoring/valuation
 # time
TlatPred <- simulTime(model=jointModel_fit, data=ND,
              raneffi=raneffi,ranefft=0, SurvCen=SurvCen)
# Predicted settlement time
Times<-mean(TlatPred,na.rm=T)
Times<-ifelse(is.na(Times),9,Times)
Times<-ifelse(Times>=9,9,Times)
True.time<-ifelse(ND$true.time[1]>=9,9,ND$true.time[1])
ND$time<-Times
raneffi<-as.numeric(RanEff[which(RanEff[,1]==uniq[i]),2])
raneff=raneffi
# Ultimate payment prediction
lfit <-predict(jointModel_fit, newdata = ND[1,],
  idVar = "id",type= "Marginal", interval= "conf", FtTimes=Times)
mean<-lfit$pred+raneff

TestRBNSPaidPredZ[i,2]<-mean
TestRBNSPaidPredZ[i,3]<- True.time
```

```
      TestRBNSPaidPredZ[i,4]<- Times}
    data.longTsub<-data.longT[!duplicated(data.longT$id, fromLast= T),]
    data.longTsub<-subset(data.longTsub,select = c(id,y))
    colnames(data.longTsub)<-c("id","y_t")
    TestRBNSPaidPredZ<-merge(TestRBNSPaidPredZ,data.longTsub, by="id")
    TestRBNSPaidPredZ$Reserve<-TestRBNSPaidPredZ$Pred-
TestRBNSPaidPredZ$y_t
    TestRBNSPaidPredZ$ActReserve<-TestRBNSPaidPredZ$yUlt-
TestRBNSPaidPredZ$y_t
    ErrorJM<-sum(TestRBNSPaidPredZ$Reserve)-
sum(TestRBNSPaidPredZ$ActReserve)

  }
} else {
  ErrorJM<-NA
}
#TotalReserve
#ErrorJM
#Error2S
#ErrorInd
```

7.3.2 Trending Techniques

For each replication, the simulated claims are evenly and independently distributed among ten accident years, and the last development year is the end of the calendar year ten. The run-off triangle is obtained by aggregating the individual claims by accident and development year. Let $C_{i,j}$ denote the cumulative payments in cell $\{i, j\}$, the the matrix for the loss triangle is given by:

$$\begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,10} \\ C_{2,1} & C_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_{10,1} & 0 & \cdots & 0 \end{bmatrix}. \tag{7.3}$$

Here, the zeros represent the unobservable cells as a result of censoring by the valuation date. Using the loss triangle matrix in (7.3), define the individual development factors as $F_{ij} = C_{i,j+1}/C_{i,j}$. Then the matrix for the F_{ij} is given by:

$$\begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & F_{1,9} \\ F_{2,1} & F_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ F_{9,1} & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 \end{bmatrix}. \tag{7.4}$$

Under a changing environment, the observed development patterns are affected by the environmental changes, which results in inaccurate predictions from the basic chain-ladder method because it relies on a stable environment. Trending techniques which are ad hoc in nature and highly dependent on actuaries' judgments are employed to treat environmental changes as a trend to adjust the development projections. To show the results can differ significantly depending on the trending algorithm used, I discuss two different algorithms based on the partitioning of the matrix for the individual development factors in (7.4) and shown in Figure 7.2.

7.3.2.1 Approach 1

For the first approach, as illustrated in the left panel of Figure 7.2, the matrix for the individual development factors F_{ij} in (7.4) is partitioned into 5 areas given by $A_1 = \{(i,j) : i \leq 5, j \leq 4\}$, $A_2 = \{(i,j) : i \leq 5, 5 \leq j \leq 10 - i\}$, $A_3 = \{(i,j) : 6 \leq i \leq 9, j \leq 10 - i\}$, $A_4 = \{(i,j) : i \leq 5, 11 - i \leq j \leq 9\}$, and $A_5 = \{(i,j) : 6 \leq i \leq 10, 11 - i \leq$

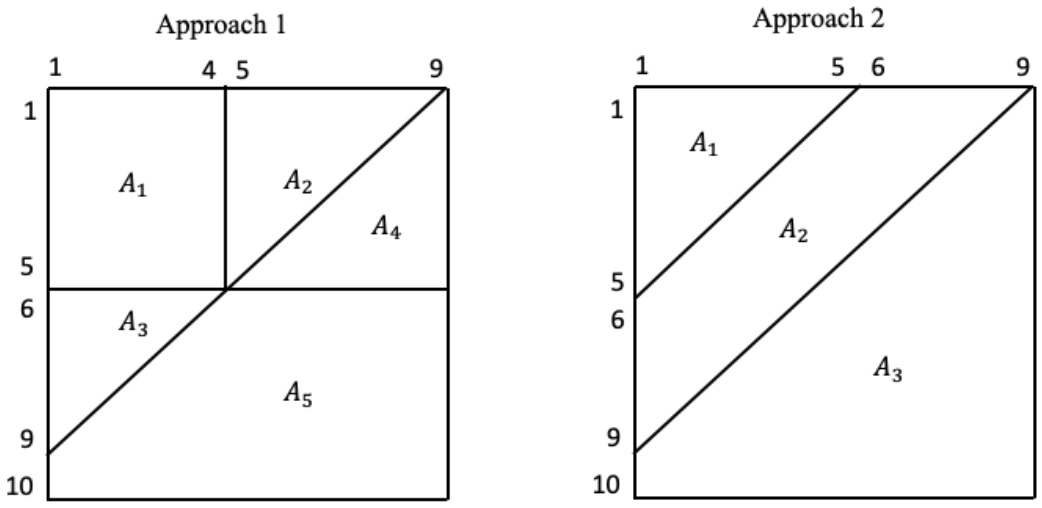

Figure 7.2: Trending algorithms based on the partitioning of matrix for the individual development factors. Each partition is defined by the columns numbers on top and row numbers on the left.

$j \leq 9\}$. Thus, the observed F_{ij} is made up of $\{A_1, A_2, A_3\}$ and the unobserved portion of F_{ij} is made up of $\{A_4, A_5\}$.

To make projections, estimates for $\{A_4, A_5\}$ needs to be obtained. The projection of the Fs in A_4 is based on the Fs in A_2. Hence, A_4 is set to equal the column averages based on A_2. It has to be noted that A_2 was not impacted by the environmental changes under the underwriting practices and policy mix scenarios but was impacted under the claims processing scenario. Then for A_4, we have:

$$F_{ij} = \bar{F}_j, \quad (i,j) \in A_4, \tag{7.5}$$

where,

$$\bar{F}_j = \sum_{i=1}^{10-j} F_{i,j}/(10-j), \quad j = 5, \ldots, 9. \tag{7.6}$$

The basic chain-ladder algorithm is applied to project the future development on A_4. To complete the projections, estimates for the partition A_5 are also needed. But due to the environmental changes, the observed historical development on A_1 and A_2 needs to be adjusted, which is done by estimating the magnitude of the impact of the environmental

changes with a rate r. Here, we have:

$$r = \frac{1}{4}\sum_{j=1}^{4}\frac{\bar{F}_j^*}{\bar{F}_j} \tag{7.7}$$

where $\bar{F}_j$, the column averages of Fs in A_1, is given by

$$\bar{F}_j = \sum_{i=1}^{5}\frac{F_{i,j}}{5}, \quad j = 1, 2, 3, 4, \tag{7.8}$$

and $\bar{F}_j^*$, the column averages of Fs in A_3, is given by

$$\bar{F}_j^* = \sum_{i=6}^{10-j}\frac{F_{i,j}}{5-j}, \quad j = 1, 2, 3, 4. \tag{7.9}$$

r is then used to make adjustments to A_1 and A_2 to take the development factors to the level after the environmental changes, shown as:

$$\bar{F}_{i,j}^* = r \cdot \bar{F}_{i,j}, \quad (i, j) \in A_1 \cup A_2. \tag{7.10}$$

With the adjusted Fs in A_1 and A_2, A_5 can be set to be the column averages of A_1, A_2 and A_3. Then, the basic chain-ladder algorithm can be applied to project the future development on A_5. This trending algorithm assumes that the actuaries are sure of the type of "trend" that they are dealing with; hence the prediction results are accurate.

7.3.2.2 Approach 2

For the second approach which is illustrated in the right panel of Figure 7.2, the matrix for the individual development factors F_{ij} in (7.4) is partitioned into 3 areas given by $A_1 = \{(i, j) : i + j \leq 6$, $A_2 = \{(i, j) : 7 \leq i + j \leq 10\}$, $A_3 = \{(i, j) : i + j \geq 11\}$. Here, the observed F_{ij} is made up of $\{A_1, A_2\}$ and the unobserved portion of F_{ij} is given by

A_3.

To make projections in A_3, only the Fs on A_2 is used. Hence, A_3 is set to equal the column averages based on A_2. We have:

$$F_{ij} = \bar{F}_j, \quad (i,j) \in A_3, \tag{7.11}$$

where

$$\bar{F}_j = \begin{cases} \sum_{i=7-j}^{10-j} \frac{F_{i,j}}{4} & \text{for} \quad j = 1, \ldots, 6, \\ \sum_{i=1}^{10-j} \frac{F_{i,j}}{10-j} & \text{for} \quad j = 7, 8, 9 \end{cases} \tag{7.12}$$

It has to be noted that this trending approach will not be accurate in adjusting the development factors affected by the environmental changes described in Section 3.4. The results using this approach illustrate that when actuaries are not sure of the type of "trend" that they are dealing with, it could lead to inaccurate predictions.

7.4 Appendix to Chapter 4

7.4.1 Estimation Results for Base Joint Model

Estimation results for the fitted joint model where y_{it} follows a Log-Normal distribution, with a Weibull baseline hazard for the survival submodel, is given in Table 7.5. All the continuous covariates in the longitudinal submodel and survival submodels are significant at a 5% significance level, except the deductible variable under the survival submodel. For the survival submodel, the association parameter $\alpha = -0.086$, though barely significant at a 5% significance level.

Table 7.5: Estimation results for base joint model: Assuming Log-Normal distribution with a linear payment trend for the longitudinal submodel and a Weibull baseline survival submodel.

Longitudinal submodel			Survival submodel		
Variable	Estimate	Std. Error	Variable	Estimate	Std. Error
(Intercept)	0.808	0.124	LnInitialEst	-0.349	0.041
TimeToPayment	0.288	0.012	LnPolicyDed	-0.003	0.012
LnInitialEst	0.853	0.010	ReportDelay	0.348	0.018
LnPolicyDed	0.028	0.008			
ReportDelay	0.029	0.014	α(association)	-0.086	0.044
Variance Components			Weibull Baseline Hazard		
σ	0.425		λ	35.449	
$\nu^{(1/2)}$	0.478		k	1.397	
Number of Payments	3,891		Number of Claims	3,264	
Categorical Variables					
Variable	LRT	df (p-value)	Variable	LRT	df (p-value)
CauseCode	90.500	9 (<0.0001)	CauseCode	95.050	9 (<0.0001)
Region	35.370	4 (<0.0001)	Region	51.500	4 (<0.0001)
EntityType	3.920	5 (0.5610)	EntityType	67.330	5 (<0.0001)
LossQtr	2.490	3 (0.4772)	LossQtr	27.420	3 (0.0001)
LossYear	17.05	3 (0.0007)	LossYear	19.95	3 (<0.0002)

7.4.2 Details for Marked Poisson Process for RBNS

Under the Marked Poisson Process framework, the likelihood for the full development process of a claim is given by (Jin, 2014):

$$L = f_V \times f_{U|v} \times f_{W|v,u} = f_V \times f_{U|v} \times f_{S|v,u} \times f_{E|v,u,s} \times f_{P|v,u,s,e}, \tag{7.13}$$

where V and U represent the claim occurrence times and reporting delay respectively. However, with the focus on RBNS reserve prediction, I am interested in the claim development process W given by

$$f_{W|v,u} = f_{S|v,u} \times f_{E|v,u,s} \times f_{P|v,u,s,e}. \tag{7.14}$$

Where S denotes the transaction occurrence times, E denotes the type of transaction, and P denotes the payment amount of the transaction. The transaction occurrence

times S are modeled by a discrete survival model with piecewise constant hazard rates. Following Jin (2014) and Antonio and Plat (2014), the first transactions are modeled with a hazard rate $g(s)$, and the later transactions are modeled with a different hazard rate $h(s)$. Let $[0, a_R]$ and $[0, b_L]$ be the interval for first and later transactions. Then we have:

$$g(s) = \sum_{r=1}^{R} g_r 1\{a_{r-1} < s \leq a_r\} \tag{7.15}$$

$$h(s) = \sum_{l=1}^{L} h_l 1\{b_{l-1} < s \leq b_l\} \tag{7.16}$$

With cumulative hazard rates given by:

$$G(s) = \int_0^s g(t)dt \tag{7.17}$$

$$H(s) = \int_0^s h(t)dt \tag{7.18}$$

Then the cumulative density functions of transaction occurrence times are given by:

$$\Pr(S_1 \leq s) = 1 - \exp(-G(s)) \tag{7.19}$$

$$\Pr(S_k \leq s) = 1 - \exp(-H(s)), \quad k > 1 \tag{7.20}$$

Let $a_R = N_1$ be regarded as the maximum waiting time to the first transaction, and $b_L = N_2$ is regarded as the maximum settlement delay. Then under these additional assumptions, the probability that the first transaction occurs at time $r, r = 1, 2, ..., N_1$ is

$$\Pr(S_1 = r | S_1 \leq N_1) = \frac{\exp\{-G(r-1)\} - \exp\{-G(r)\}}{1 - \exp\{-G(N_1)\}} \tag{7.21}$$

And given the occurrence time of the first transaction, $S_{k-1} = c$, the probability that transaction k occurs at time $r, r = c+1, c+2, ..., N_2$ is

$$\Pr(S_k = r | S_{k-1} = c, S_k \leq N2) = \frac{\exp\{-H(r-1)\} - \exp\{-H(r)\}}{\exp\{-H(c)\} - \exp\{-H(N_2)\}} \tag{7.22}$$

For the Wisconsin LGPIF training dataset, the maximum waiting time for the first transaction is 17 months, and the maximum settlement delay is 27 months. It is assumed that there is at most one transaction in each month, and the transactions can only occur at the end of a month. As noted in Jin (2014), this discrete setup is consistent with the fact that many insurers aggregate transactions on a monthly basis by the end of each month. Therefore, the piecewise-constant hazard rates is defined to have jumps every month, i.e. $a_1 = 0, a_2 = 1, \cdots, a_{17} = 17$ and $b_1 = 0, b_2 = 1, \cdots, b_{27} = 27$.

Furthermore, for the type of transactions E, I consider two types for claim i at time $S = s$; a payment transaction that leads to settlement ($e_{is} = 1$) and an intermediate payment transaction ($e_{is} = 0$). With an intermediate transaction, the claim development process continues. Given a transaction at time s, the transaction type is determined by a logit model that accommodates heterogeneity by incorporating random effects a_i.The probabilities also depend on the time of the transaction and covariates x_{is} given by:

$$\Pr(e_{is} = 1 | a_i) = \pi(x'_{is}\beta + a_i) = \frac{1}{1 + \exp(-(x'_{is}\beta + a_i))}. \tag{7.23}$$

To model the incremental payments P, a Generalized Linear Mixed-Effects Model is specified.

7.5 Appendix to Chapter 5

7.5.1 Summary statistics at the policy and claim level

Table 7.6 summarizes the distribution of the claim frequency and severity, which are the two continuous outcomes of interest at the policy level. The Spearman correlation coefficient shows that the two outcomes are correlated. The table also shows that the coverage and deductible distributions are right-skewed, and I deal with the skewness by taking logarithmic transformations on these variables. The high correlation coefficient for coverage suggests it will be a significant predictor in the claim frequency and severity model. In addition, the table summarizes the transaction frequency to the settlement of a reported claim and the average payment amounts; the Spearman correlation coefficient shows that they are correlated. Moreover, the correlation coefficient between deductible and the transaction frequency and severity indicates it may be used as a predictor.

Table 7.6: Summary statistics for outcomes at the policy level (claim frequency and severity) and claim level (transaction frequency and severity), and continuous covariates (deductibles, and coverages).

	Policy Level					
	Min.	Median	Average	Max.	Claim Frequency (ρ_S)	Claim Severity (ρ_S)
Claim Frequency	0	0	0.893	231	-	0.964
Claim Severity	0	0	9,649	12,922,218	0.964	-
Deductible	500	1,000	3,407	100,000	0.051	0.090
Coverage(000'S)	0.4	11,493	11,493	2,444,797	0.404	0.396
	Claim Level					
	Min.	Median	Average	Max.	Transaction Frequency (ρ_S)	Transaction Payment (ρ_S)
Transaction Frequency	1	1	1.221	11	-	0.301
Transaction Payment	8.2	2,609	11,757	1,174,293	0.301	-
Deductible	500	1,000	6,739	100,000	0.095	0.132
Coverage(000'S)	102.3	58,046	295,660	2,444,797	-0.019	-0.079

Table 7.7 shows the summary statistics of the claim frequency and severity at the policy level and the transaction frequency and severity at the claim level for the categorical variables in the dataset. The table suggests a high variation in the claim frequency and

average severity of the claims across the categorical variables at the policy level. At the claim level, the transaction frequency does not vary much, but there is a substantial variation with the average severity.

Table 7.7: Summary statistics at the policy and claim level by categorical variables.

	Policy Level		Claim Level	
Variable	Average Frequency	Average Severity	Average Frequency	Average Severity
Entity Type				
Village	0.349	2,977	1.196	7,776
City	1.607	12,015	1.222	9,276
County	3.330	17,453	1.169	13,089
Misc	0.172	4,110	1.127	14,328
School	0.931	25,963	1.231	15,604
Town	0.092	1,204	1.126	6,351
Region				
Northeastern	0.561	7,835	1.189	12,840
Northern	0.410	10,682	1.229	13,879
Southeastern	1.389	30,123	1.206	15,605
Southern	1.196	5,820	1.263	7,921
Western	0.483	5,210	1.216	7,989
Alarm Credit				
No Alarm Credit	0.226	2,427	1.237	7,873
Alarm Credit 5%	0.290	3,508	1.183	10,566
Alarm Credit 10%	0.275	3,016	1.165	10,250
Alarm Credit 15%	1.059	15,869	1.212	11,326
Alarm Credit (Combination)	2.212	29,923	1.233	14,476

Bibliography

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

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

ASTIN (2016). Non-life reserving practices. *ASTIN working party report*.

ASTIN (2017). Individual claim development with machine learning. *ASTIN working party report*.

Avanzi, B., Wong, B., and Yang, X. (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance Mathematics and Economics*, 71:1–14.

Badescu, A. L., Lin, X. S., and Tang, D. (2016a). A marked cox model for the number of ibnr claims: Estimation and application. *SSRN*.

Badescu, A. L., Lin, X. S., and Tang, D. (2016b). A marked cox model for the number of ibnr claims: Theory. *Insurance Mathematics and Economics*, 69:29–37.

Baudry, M. and Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 35(5):1127–1155.

Berquist, J. R. and Sherman, R. E. (1977). Loss reserve adequacy testing: A comprehensive systematic approach. *Proceedings of the Casualty Actuarial Society, LXVII*, pages 123–184.

Bornhuetter, R. . L. and Ferguson, R. (1972). The actuary and ibnr. *CAS*, LIX:181–195.

Carrato, A., Concina, F., Gesmann, M., Murphy, D., Wüthrich, M., and Zhang, W. (2020). Claims reserving with r: Chainladder-0.2.11 package vignette. *https://cran.r-project.org/web/packages/ChainLadder/vignettes/ChainLadder.pdf*.

Chi, Y.-Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62:432–445.

Cook, R. J. and Lawless, J. F. (2007). *The statistical analysis of recurrent events*. Springer.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Crevecoeur, J., Antonio, K., and Verbelen, R. (2019). Modeling the number of hidden events subject to observation delay. *European Journal of Operational Research*, 277(3):930–944.

Cummins, J. D. (2002). Property-liability insurance price deregulation: the last bastion? In Cummins, J. D., editor, *Deregulating Property-Liability Insurance:Restoring Competition and Increasing Market Efficiency*, chapter 1, pages 1–24. Brooking Institution Press, Washington DC.

De Felice, M. and Moriconi, F. (2019). Claim watching and individual claims reserving using classification and regression trees. *Risks*, 7(4).

Dean, C. G. (2014). Generalized linear models. In Frees, E. W., Derrig, R. A., and Meyers, G., editors, *Predictive Modeling Applications in Actuarial Science*, chapter 5, pages 107–137. Cambridge, MA: Cambridge University Press.

DeGruttola, V. and Tu, X. (1994). Modeling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014.

Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 43:49–73.

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

Elashoff, R. M., Li, G., and Li, N. (2016). *Joint Modeling of Longitudinal and Time-to-Event Data*. Chapman and Hall/CRC.

England, P. D. and Verrall, R. J. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance Mathematics and Economics*, 25:281–293.

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

Frees, E. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. New York: Cambridge University Press.

Frees, E. W. (2014). Frequency and severity models. In Frees, E. W., Derrig, R. A., and Meyers, G., editors, *Predictive Modeling Applications in Actuarial Science*, chapter 6, pages 138–164. Cambridge, MA: Cambridge University Press.

Frees, E. W., Meyers, G., and Cummings, D. (2011). Summarizing insurance scores using a gini index. *Journal of the American Statistical Association*, 106:1085–1098.

Frees, E. W., Shi, P., and Valdez, E. A. (2009). Actuarial applications of a hierarchical insurance claims model. *Astin Bulletin*, 39(1):165–197.

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

Friedland, J. (2010). *Estimating Unpaid Claims Using Basic Techniques*. Casualty Actuarial Society.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

Guszcza, J. and Lommele, J. (2006). Loss reserving using claim-level data. *In Annual meeting for the Casualty Actuarial Society, November 2006*.

Hachemeister, C. and Stanard, J. (1975). Ibnr claims count estimation with static lag functions. *Spring Meeting of the Casualty Actuarial Society*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical learning 2nd ed.* New York: Springer.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492.

Henckaerts, R., Antonio, K., Clijsters, M., and Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, 2018(8):681–705.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796–2801.

Jewell, W. S. (1989). Predicting ibnyr events and delays, part i continuous time. *ASTIN Bulletin*, 19(1):25–55.

Jin, X. (2014). Micro-level stochastic loss reserving models for insurance. *The University of Wisconsin-Madison, ProQuest Dissertations Publishing*.

Kahn, J. (2002). Reserving for runoff operations – a real life claims specific methodology for reserving a workers' compensation runoff entity. *Casualty Actuarial Society Forum*, pages 139–210.

Kremer, E. (1982). Ibnr-claims and the two-way model of anova. *Scandinavian Actuarial Journal*, 1982(1):47–55.

Kuang, D., Nielsen, B., and Nielsen, J. (2008). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95(4):979–986.

Kuang, D., Nielsen, B., and Nielsen, J. (2011). Forecasting in an extended chain-ladder-type model. *Journal of Risk and Insurance*, 78(2):345–359.

Lange, K. (2004). *Optimization*. Springer-Verlag, New York.

Larsen, C. R. (2007). An individual claims reserving model. *ASTIN Bulletin*, 37(1):113–132.

Little, Roderick, J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.

Little, R. (2008). Selection and pattern-mixture models. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, chapter 18, pages 409–432. CRC Press, Boca Raton.

Liu, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*, 28(6):972–986.

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

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

Martínez-Miranda, M. D., Nielsen, B., Nielsen, J. P., and Verrall, R. (2011). Cash flow simulation for a model of outstanding liabilities based on claim amounts and claim numbers. *ASTIN Bulletin*, 41(1):107–129.

Martínez-Miranda, M. D., Nielsen, J. P., Sperlich, S., and Verrall, R. (2013). Continuous chain ladder:reformulating and generalizing a classical insurance problem. *Expert Systems with Applications*.

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

Mikosch, T. (2009). *Non-Life Insurance Mathematics; An Introduction with the Poisson Process*. Springer-Verlag.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data In: Springer Series in Statistics*. Springer, New York.

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

Norberg, R. (1999). Prediction of outstanding liabilities ii. model variations and extensions. *ASTIN Bulletin*, 29(1):5–25.

Parodi, P. (2012). Triangle-free reserving: A non-traditional protocol for estimating reserves and reserve uncertainty. *In Proceedings of GIRO*.

Petroni, K. R. (1992). Optimistic reporting in the property-casualty insurance industry. *Journal of Accounting and Economics*, 15(4):485–508.

Poon, J. H. L. (2019). Penalising unexplainability in neural networks for predicting payments per claim incurred. *Risks*, 7(3).

Renshaw, A. E. and Verrall, R. J. (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4(04):903–923.

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*, 35(9):1–33.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC.

Rizopoulos, D. (2016). The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(1):1–46.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society,Series B (Statistical Methodology)*, 71(3):637–654.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74.

Shi, P., Feng, X., and Boucher, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics*, 10(2):834–863.

Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58(4):742–753.

Sweeting, M. J. and Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763.

Taylor, G. (2014). Claims triangles/loss reserves. In Frees, E. W., Derrig, R. A., and Meyers, G., editors, *Predictive Modeling Applications in Actuarial Science*, chapter 18, pages 449–480. Cambridge, MA: Cambridge University Press.

Taylor, G. and Campbell, M. (2002). Statistical case estimation. *In Research paper number 104, The University of Melbourne, Australia.*

Taylor, G. C. and McGuire, G. (2004). Loss reserving with glms: A case study. *In Annual Meeting for the Casualty Actuarial Society, Spring 2004.*

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

Thorne, J. O. (1978). Loss reserve adequacy testing: A comprehensive systematic approach [discussion]. *Proceedings of the Casualty Actuarial Society, LXV*, pages 10–33.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14(3):809–834.

Tsiatis, A., DeGruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error: applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90:27–37.

Verbeke, G., Molenberghs, G., and Rizopoulos, D. (2010). Random effects models for longitudinal data. In van Montfort, K., Oud, J. H., and Satorra, A., editors,

Longitudinal Research with Latent Variables, chapter 2, pages 37–96. Springer, Berlin, Heidelberg, Berlin.

Verrall, R., Nielsen, J. P., and Jessen, A. H. (2010). Prediction of rbns and ibnr claims using claim amounts and claim counts. *ASTIN Bulletin*, 40(2):871.

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

Verrall, R. J. and Wüthrich, M. V. (2016). Understanding reporting delay in general insurance. *Risk*.

Werner, G. and Modlin, C. (2016). *Basic Ratemaking, Fifth Edition*. Casualty Actuarial Society.

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

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

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

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

Wüthrich, M. V. and Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. John Wiley & Sons.

Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-curve models and their application to prostate cancer. *Statistica Sinica*, 14(3):835–862.