JOINT MODEL PREDICTION AND APPLICATION TO INDIVIDUAL-LEVEL LOSS RESERVING

BY

A. NII-ARMAH OKINE, EDWARD W. FREES AND PENG SHI

ABSTRACT

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, we 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 claim settlement. We discuss statistical inference and focus on the prediction aspects of the model. We demonstrate applications of the proposed model in the reserving practice with a detailed empirical analysis using data from a property insurance provider. The prediction results from an out-of-sample validation show that the joint model framework outperforms existing reserving models that ignore the payment–settlement association.

KEYWORDS

Dynamic prediction, joint model for longitudinal and survival data, micro-level loss reserving, RBNS reserves.

JEL codes: C33, C51, C52, C53, C55, G22.

1. INTRODUCTION

A loss reserve represents the insurer's best estimate of outstanding liabilities for claims that occurred on or before a valuation date. Inaccurate prediction of unpaid claims 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 instance, inadequate reserves could lead

Astin Bulletin 52(1), 91–116. doi:10.1017/asb.2021.28

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.

In claim 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. See Figure 3 for an example in property insurance.

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, 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 would 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 are negatively correlated.

Further, the payment–settlement association suggests 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 in the sense, when more information becomes available over time, an actuary could use claim history to update the prediction for the settlement time and ultimate claim payments.

The goal of this paper 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 claim prediction. In doing so, we employ a joint modeling framework developed in the statistical literature for longitudinal outcomes and time-to-event data. The joint model (JM) for reserving purposes consists of two submodels, the longitudinal submodel governs the payment process for a given claim, and the survival submodel 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 affect the instantaneous settlement probability, and the settlement intensity determines whether there are further payments.

For statistical inference of the joint model, we discuss both estimation and prediction with the focus on the latter. The properties of estimators and predictions are investigated using simulation studies, and we find that the advantages of the proposed joint model are more pronounced for long-tail lines of business. Furthermore, we present a detailed empirical analysis of the joint model

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

framework using data from a property insurance provider with the focus on the Reported But Not Settled (RBNS) reserve prediction. We fit thejoint model to a training data set and find significant positive relation between the payment history and settlement time. The RBNS prediction performance of the joint model is compared to existing reserving models using out–of–sample data, and the results suggest that accounting for the payment–settlement association leads to better prediction.

We contribute to the literature in the following aspects: First, we introduce the joint model for longitudinal and time-to-event data into the micro-level loss reserving literature, and thus provide a novel solution to the sample selection issue that is due to the association between the size of claims and time of settlement. Second, because of the predictive nature of loss reserving, we investigate the predictive performance of the joint model, using both simulated and real-world data, which enriches the existing statistical literature that has primarily focused on the estimation aspect of inference. Third, the detailed analysis not only provides empirical evidence of the payment–settlement association and its role in the dynamic prediction but also provides guidance for practitioners to employ the proposed method in practice.

The rest of the paper is organized as follows: Section 2 reviews the literature on current loss reserving methods and joint models for longitudinal and time-to-event data. Section 3 introduces the joint modeling framework for individual-level loss reserving. Section 4 discusses estimation and prediction for the joint model. Section 5 evaluates the properties of the model using simulation studies. Section 6 describes the property insurance claims data set and its important characteristics that motivate the joint modeling framework, provides estimation results using a training data set and prediction results using a hold-out sample, and discusses limitations. Section 7 concludes the paper.

2. Literature review

2.1. Literature on reserving models

In the actuarial literature, there are two main classes of reserving techniques: macro-level and micro-level. The macro-level models are based on aggregate claims data summarized in run-off triangles, and the reserve is estimated using the chain-ladder (CL) method and its extensions. See Wüthrich and Merz (2008) for a comprehensive review on macro-level reserving methods. The ease of implementation and interpretation is a major strength of macro-level models, but they come with a risk of inaccurate predictions. 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. In case of environmental changes, some of the commonly-used macro-models can generate a reserve estimate without material errors. In

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

this context, "environmental changes" refers to changes in the insurer's business that can affect loss reserving, for example, underwriting practices, claims processing, mix of products, and so forth. To handle environmental changes, macro-level methods consider either expected claims that allow actuaries to incorporate a priori reserve estimate, or trending techniques that treat environmental change as trend to adjust the development projections. However, highly dependent on actuaries' judgments, both techniques could lead to problematic reserve estimates (Jin, 2014).

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 claim development in the prediction of 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. The most studied method is the marked Poisson process (MPP) framework introduced by Arjas (1989), Jewell (1989), Norberg (1993) and (1999). Antonio and Plat (2014) provided the first empirical study with data from a personal-line general liability insurance portfolio. The 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. The marked Cox process provides an extension to allow for overdispersion and serial dependence (Avanzi *et al*., 2016; Badescu *et al*., 2016b and 2016a). 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 claims (Taylor and Campbell, 2002; Taylor and McGuire, 2004 and Taylor *et al*., 2008). Most recently, machine learning algorithms have become popular in individual-level loss reserving because they are highly flexible and can deal with structured and unstructured information (for example, see Wüthrich, 2018a and 2018b).

2.2. Literature on joint models

The existing micro-level reserving methods do not explicitly capture the dependence between the payment history and settlement process. Recently, papers such as Lopez *et al*., (2019) attempted to handle the bias due to the payment–settlement association using a weighting scheme. But their framework does not explicitly capture the dependence between the payment history and settlement process. We further extend the literature by introducing the joint longitudinal survival model framework to handle such association explicitly.

The joint model has been proposed in the medical statistics literature for modeling longitudinal and survival outcomes when the two components are

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

correlated (Elashoff *et al*., 2017). Two general frameworks have received extensive attention, the pattern mixture model and the selection model (Little, 2008). These two frameworks differ in the way the joint distribution is factorized. In the former, the joint distribution is specified using the marginal distribution of time-to-event outcome and the conditional distribution of longitudinal outcomes given the time-to-event outcome. In contrast, the joint distribution in the latter is specified using models for the marginal distribution of longitudinal outcomes and the conditional distribution of time-to-event outcome given longitudinal outcomes. Diggle and Kenward (1994) were the first to apply selection models to nonrandom 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. The two model families are primarily applied with discrete drop out times and cannot be easily extended to continuous time.

The properties of the joint models have been well developed in the biomedical literature in clinical studies (Ibrahim *et al*., 2010) and non-clinical studies (Liu, 2009). Tsiatis and Davidian (2004), Yu *et al*., (2004), and Verbeke *et al*., (2010) give excellent overviews of joint models. Besides, Rizopoulos (2010) and (2016) develop R packages for joint models.

3. Joint model for claim payment and settlement

3.1. General framework

In this section, we introduce the joint model framework to the loss reserving problem, focusing on a subset of selection models called shared-parameter models. In shared-parameter models, a latent random effects $\boldsymbol{b}_i$ is used to capture the association between the longitudinal and the time-to-event outcomes (Rizopoulos, 2012). For this application, 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, we set the time origin for a claim as its reporting time. For the ith claim ($i = 1, \ldots, N$), we 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 $\Delta_i = I(T_i^* < c_i)$, such that $I(A) = 1$ 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 $\{Y_i(t):0 \le t \le T_i^*\}$ be the payment process, and $\boldsymbol{Y}_i^* = \{Y_{it}, t \in \tau_i^*\}$ be the vector of the realized 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, we define $\tau_i = \{t_{ij}; j = 1, \ldots, n_i\}$ as the observable payment times and denote

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

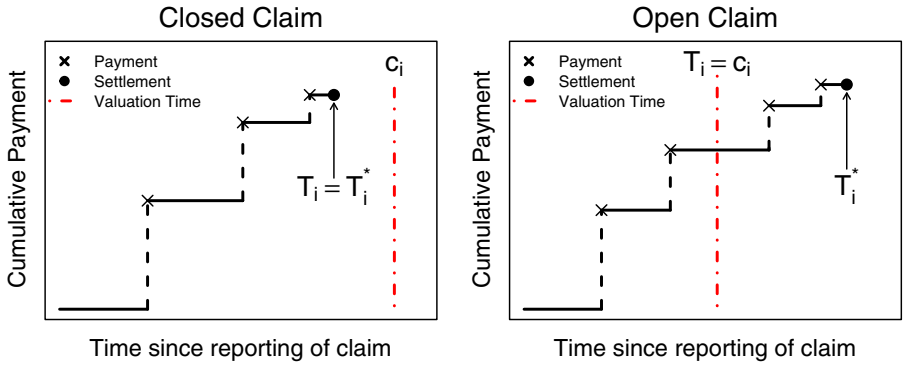

FIGURE 1: Graphical illustration of the cumulative payment process from the time of reporting to settlement.

$\boldsymbol{Y}_i = \{Y_{it}, t \in \tau_i\}$ the vector of cumulative payments at observed time of payments. Further denote $\boldsymbol{Y}_i^+ = \{Y_{it}, t \in \tau_i^+\}$ the vector of cumulative payments at future times $\tau_i^+ = \{t_{ij}; j = n_i + 1, \ldots, n_i^*\}$ after the valuation time. In the joint model framework, the joint distribution of $f_{\boldsymbol{Y}_i^*, T_i^*}(\boldsymbol{y}_i^*, t_i^*)$ is defined by

$$f_{\boldsymbol{Y}_i^*, T_i^*}(\boldsymbol{y}_i^*, t_i^*) = \int f(\boldsymbol{y}_i^* | \boldsymbol{b}_i) f(t_i^* | \boldsymbol{b}_i) dF(\boldsymbol{b}_i), \tag{3.1}$$

where $\boldsymbol{b}_i$ denotes the vector of random effects that account for the claim-specific unobserved heterogeneity. The formulation in (3.1) relies on a conditional independence assumption. Figure 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 size of the jump represents the amount of incremental payment. The left panel presents a closed claim where the entire development process of the claim is observed before the valuation time, that is $(\Delta_i = 1, n_i = n_i^*)$. 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, that is $(\Delta_i = 0, n_i \leq n_i^*)$.

The rest of the section focuses on the general modeling framework for both the longitudinal and survival submodels. Alternative model specifications under the general framework are considered in the empirical analysis and thus discussed in Section 6.

3.2. Longitudinal submodel of claim payments

The cumulative payments Y_{it} are specified using generalized linear mixed effect models (GLMM). See, for instance, Frees (2004) and Molenberghs and Verbeke (2006) for details. Then, conditional on the random effects $\boldsymbol{b}_i$, the cumulative payment Y_{it} is assumed to be from the exponential family with dispersion parameter ϕ. The conditional mean of Y_{it}, $\mu_{it} = E[Y_{it} | \boldsymbol{b}_i]$, is specified

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

as a linear combination of covariates via a link function $g(\cdot)$, that is

$$\eta_{it} = g(\mu_{it}) = f(t_i; \boldsymbol{\beta}_1) + \boldsymbol{x}'_{it}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{it}\boldsymbol{b}_i. \tag{3.2}$$

Here, $f(t_i;\boldsymbol{\beta}_1)$ is a function of payment time t parameterized by $\boldsymbol{\beta}_1$. Examples are linear functions, polynomial functions, and splines. Furthermore, $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ are the vectors of covariates in the fixed and random effects, respectively, and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ is the regression coefficients to be estimated. In addition, $\boldsymbol{b}_i$, $i = 1, \ldots, N$, are independent of each other and $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is the covariance matrix with unknown parameters $\boldsymbol{v}$. We emphasize that the covariates are indexed by t because, in addition to time-independent covariates like claim codes, the model allows us to include time-dependent covariates. In this model, we assume Y_{it} are independent across time conditional on random effects $\boldsymbol{b}_i$ and predictors $\boldsymbol{x}_{it}$.

3.3. 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

$$h_i(t|\eta_{it}) = h_0(t)\exp\{\boldsymbol{w}'_{it}\boldsymbol{\gamma} + \alpha\eta_{it}\}, \tag{3.3}$$

where $h_0(t)$ is the baseline hazard, $\boldsymbol{w}_{it}$ is a vector of covariates, and $\boldsymbol{\gamma}$ is the corresponding regression coefficients. 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, that is larger payments will accelerate the settlement, and vice versa. From (3.3), the survival function of T_i^* is

$$S_i(t|\eta_{it}) = \exp\left(-\int_0^t h_0(s)\exp\{\boldsymbol{w}'_{is}\boldsymbol{\gamma} + \alpha\eta_{is}\}ds\right). \tag{3.4}$$

For the baseline hazard in (3.3), we consider both the Weibull model and an approximation based on splines. The Weibull baseline is given by

$$h_0(t) = \lambda\kappa t^{\kappa-1}, \tag{3.5}$$

where λ is the scale parameter, and κ is the shape parameter. When $\kappa = 1$, $h_0(t)$ reduces to an exponential baseline function. The Weibull model is commonly used because of its simplicity and the easy interpretability. A more flexible model is to approximate the baseline hazard using splines. Specifically, we consider:

$$\log h_0(t) = \lambda_0 + \sum_{k=1}^{K} \lambda_k B_k(t, q). \tag{3.6}$$

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 $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \cdots, \lambda_K)$ are the spline coefficients. For convenience, we denote $\boldsymbol{\omega}$ to be

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

the parameters in the baseline hazard model. Then, $\boldsymbol{\omega} = \{\kappa, \lambda\}$ for the Weibull baseline and $\boldsymbol{\omega} = \{\lambda_0, \lambda_1, \cdots, \lambda_K\}$ for the spline baseline.

4. Statistical Inference

4.1. Estimation

The parameters of the joint model are estimated using likelihood-based methods. Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 = \{\boldsymbol{\beta}, \boldsymbol{\nu}, \phi\}$ summarizes the parameters of the longitudinal submodel including both regression coefficients and variance components, and $\boldsymbol{\theta}_2 = \{\boldsymbol{\omega}, \boldsymbol{\gamma}, \alpha\}$ summarizes the parameters of the survival submodel that includes baseline hazard, regression coefficients, and association between claim payment and settlement. The likelihood function for observables $(t_i, \delta_i, \boldsymbol{y}_i)$, that are based on random variables $(T_i, \Delta_i, \boldsymbol{Y}_i)$, of claim i is shown as

$$\begin{aligned} L(\boldsymbol{\theta}; t_i, \delta_i, \boldsymbol{y}_i) &= \int f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}) f(t_i, \delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}) dF(\boldsymbol{b}_i;\boldsymbol{\theta}). \\ &= \int \left[\prod_{t\in\tau_i} f(y_{it}|\boldsymbol{b}_i;\boldsymbol{\theta})\right] f(t_i, \delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}) f(\boldsymbol{b}_i;\boldsymbol{\theta}) d\boldsymbol{b}_i, \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} f(t_i, \delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}) &= (h_i(t_i|\boldsymbol{b}_i;\boldsymbol{\theta}))^{\delta_i} S_i(t_i|\boldsymbol{b}_i;\boldsymbol{\theta}) \\ &= \left(h_0(t_i) \exp\{\boldsymbol{w}'_{it_i} \boldsymbol{\gamma} + \alpha \eta_{it_i}\}\right)^{\delta_i} \exp\left(-\int_0^{t_i} h_0(s) \exp\{\boldsymbol{w}'_{is} \boldsymbol{\gamma} + \alpha \eta_{is} ds\}\right). \end{aligned} \tag{4.2}$$

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

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log L(\boldsymbol{\theta}; t_i, \delta_i, \boldsymbol{y}_i). \tag{4.3}$$

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

$$I(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{N} \frac{\partial^2 \log L(\boldsymbol{\theta}; t_i, \delta_i, \boldsymbol{y}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}, \tag{4.4}$$

and the second-order derivative is approximated by the numerical Hessian matrix. Song *et al*., (2002) proposed an estimation procedure that does not require normality assumption for random effects and showed that estimation under normal assumption is robust to misspecification. In addition, Rizopoulos *et al*., (2008) showed that misspecification of the random effects

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

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 (4.1) and the integral in the survival density function (4.2). 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 paper, we employ the Gaussian quadrature numerical techniques to evaluate the likelihood function. It is worth noting that the computational aspect of the model is not the focus of our study. We refer to the aforementioned literature for details.

The random-effects estimate $\hat{\boldsymbol{b}}_i$ for claim specific predictions is obtained using Bayesian methods with posterior distribution:

$$f(\boldsymbol{b}_i|t_i, \delta_i, \boldsymbol{y}_i;\hat{\boldsymbol{\theta}}) = \frac{f(t_i, \delta_i|\boldsymbol{b}_i;\hat{\boldsymbol{\theta}})f(\boldsymbol{y}_i|\boldsymbol{b}_i;\hat{\boldsymbol{\theta}})f(\boldsymbol{b}_i;\hat{\boldsymbol{\theta}})}{f(t_i, \delta_i, \boldsymbol{y}_i;\hat{\boldsymbol{\theta}})}. \tag{4.5}$$

The mean $\hat{\boldsymbol{b}}_i$ of the posterior distribution is used as the empirical Bayes estimate and is obtained by

$$\hat{\boldsymbol{b}}_i = \int \boldsymbol{b}_i f(\boldsymbol{b}_i|t_i, \delta_i, \boldsymbol{y}_i;\hat{\boldsymbol{\theta}}) d\boldsymbol{b}_i. \tag{4.6}$$

4.2. Prediction

At the valuation time, an open claim i is characterized by the time since reporting c_i and longitudinal claim history $\mathcal{Y}_i(c_i) = \{y_{it}, 0 \le t \le c_i\}$. Since the claim is open, it implies that the settlement time $T_i^* > c_i$. With the fitted joint model, we obtain the RBNS reserve prediction for the ith claim at the valuation time, $\hat{R}_i^{RBNS}(c_i)$, 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 (4.11) from Section 4.2.1.

(b) Predict the ultimate payment, $\hat{Y}_i^{ULT}(u)$, given $\mathcal{Y}_i(c_i)$ using (4.12) from Section 4.2.2.

(c) With the cumulative payment for the ith claim at valuation time, $Y_i(c_i)$, we have:

$$\hat{R}_i^{RBNS}(c_i) = \hat{Y}_i^{ULT}(u) - Y_i(c_i). \tag{4.7}$$

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

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

$$\hat{R}^{RBNS}(c) = \sum_{i=1}^{m} \hat{R}_i^{RBNS}(c_i). \tag{4.8}$$

4.2.1. 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:

$$\pi_i(u|c_i) = \Pr\left(T_i^* \geq u | T_i^* > c_i, \mathcal{Y}_i(c_i); \boldsymbol{\theta}\right) = \frac{S_i\left(u|\eta_{iu}, \boldsymbol{w}_{iu}; \boldsymbol{\theta}\right)}{S_i\left(c_i|\eta_{ic_i}, \boldsymbol{w}_{ic_i}; \boldsymbol{\theta}\right)}, \tag{4.9}$$

where $S_i(\cdot)$ is given by (3.4), and $u > c_i$. $\boldsymbol{w}_{iu}$ and $\boldsymbol{w}_{ic_i}$ 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 $\hat{\boldsymbol{\theta}}$ and the empirical Bayes estimate $\hat{\boldsymbol{b}}_i$, an estimate of $\pi_i(u|c_i)$ is given by

$$\hat{\pi}_i(u|c_i) = \frac{\hat{S}_i\left(u|\hat{\eta}_{iu}, \boldsymbol{w}_{iu}; \hat{\boldsymbol{\theta}}\right)}{\hat{S}_i\left(c_i|\hat{\eta}_{ic_i}, \boldsymbol{w}_{ic_i}; \hat{\boldsymbol{\theta}}\right)}, \tag{4.10}$$

where $\hat{\eta}_{iu} = f(u; \hat{\boldsymbol{\beta}}_1) + \boldsymbol{x}'_{iu}\hat{\boldsymbol{\beta}}_2 + \boldsymbol{z}'_{iu}\hat{\boldsymbol{b}}_i$ and $\hat{\eta}_{ic_i} = f(c_i; \hat{\boldsymbol{\beta}}_1) + \boldsymbol{x}'_{ic_i}\hat{\boldsymbol{\beta}}_2 + \boldsymbol{z}'_{ic_i}\hat{\boldsymbol{b}}_i$. 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

$$\hat{u}_i = \int_{c_i}^{\infty} \hat{\pi}_i(u|c_i) du. \tag{4.11}$$

4.2.2. 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

$$\hat{Y}_i(u) = g^{-1}(f(u; \hat{\boldsymbol{\beta}}_1) + \boldsymbol{x}'_{iu}\hat{\boldsymbol{\beta}}_2 + \boldsymbol{z}'_{iu}\hat{\boldsymbol{b}}_i). \tag{4.12}$$

Here, $g^{-1}(\cdot)$ is the inverse of the link function, $\{\boldsymbol{x}_{iu}, \boldsymbol{z}_{iu}\}$ are covariates, and $\hat{\boldsymbol{\beta}} = \{\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\}$ are the maximum likelihood estimates. The point prediction of the ultimate amount of claim i, $\hat{Y}_i^{ULT}(u)$, is given by the mean of all expected cumulative payments at simulated future times $u > c_i$.

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

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

Algorithm 1 Reserve prediction routine for joint model.

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

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

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

  Calculate $\hat{\eta}_{ic_i} = f(c_i; \hat{\boldsymbol{\beta}}_1) + \boldsymbol{x}'_{ic_i}\,\hat{\boldsymbol{\beta}}_2 + \boldsymbol{z}'_{ic_i}\,\hat{\boldsymbol{b}}_i$;

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

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

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

    Calculate $u_{ik} = \hat{H}_i^{-1}(-\log(U \times \hat{S}_i(c_i|\hat{\eta}_{ic_i})))$;

    where $\hat{H}_i(u) = \int_0^u \hat{h}_0(s)\exp\{\boldsymbol{w}'_{is}\,\hat{\boldsymbol{\gamma}} + \hat{\alpha}\hat{\eta}_{is}\}ds$;

  end for

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

  Generate $\hat{Y}_{ik}(u_{ik}) = g^{-1}(\hat{\eta}_{iu_{ik}})$; $\hat{\eta}_{iu_{ik}} = f(u_{ik}; \hat{\boldsymbol{\beta}}_1) + \boldsymbol{x}'_{iu_{ik}}\,\hat{\boldsymbol{\beta}}_2 + \boldsymbol{z}'_{iu_{ik}}\,\hat{\boldsymbol{b}}_i$;

  Calculate $\hat{Y}_i^{ULT}(u) = K^{-1}\sum_{k=1}^{K} \hat{Y}_{ik}(u_{ik})$; For ultimate amount point prediction.

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

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

end for

5. Estimation and prediction performance evaluation using simulated data

To better understand the strength and limitations of the proposed joint model, we investigate the performance of estimation and prediction routines described in Section 4 using simulated data.

5.1. 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

$$\eta_{it} = g(E[Y_{it}|\boldsymbol{b}_i]) = f(t_i; \boldsymbol{\beta}_1) + \boldsymbol{x}'_{it}\,\boldsymbol{\beta}_2 + \boldsymbol{z}'_{it}\,\boldsymbol{b}_i = \beta_{10} + t\beta_{11} + x_{i1}\beta_{21} + x_{i2}\beta_{22} + b_{i0}, \tag{5.1}$$

where $f(t_i; \boldsymbol{\beta}_1) = \beta_{10} + t\beta_{11}$ is a linear function of the payment times, and $\boldsymbol{x}_{it} = \{x_{i1}, x_{i2}\}$. The random effects are generated from a normal distribution with mean zero and variance ν, $\mathcal{N}(0, \nu)$. The survival submodel is a proportional hazards model with an exponential base hazard. Specifically, with $\boldsymbol{w}_{it} = \{x_{i1}, x_{i2}\}$, the conditional hazard function is

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

TABLE 1.

ESTIMATION RESULTS FOR JOINT MODEL 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)									
$\beta_{10} = 1.0$	0.003	0.001	−0.008	0.005	0.005	0.004	0.059	0.056	0.051
$\beta_{11} = 0.3$	0.001	0.002	0.001	0.001	0.001	0.001	0.011	0.010	0.010
$\beta_{21} = 0.2$	−0.008	−0.001	0.001	0.004	0.003	0.004	0.053	0.039	0.042
$\beta_{22} = 0.4$	−0.002	−0.002	0.006	0.004	0.003	0.003	0.044	0.042	0.039
$\nu = 0.09$	0.000	−0.001	0.000	0.001	0.001	0.001	0.018	0.015	0.016
$\sigma = 1.5$	0.004	0.001	0.005	0.005	0.004	0.003	0.055	0.043	0.038
Survival submodel									
$\gamma_1 = 0.5$	0.000	−0.004	0.000	0.008	0.007	0.007	0.101	0.085	0.081
$\gamma_2 = 0.3$	0.007	−0.001	0.000	0.009	0.007	0.006	0.106	0.079	0.078
$\log(\lambda) = -1.139$	−0.036	−0.021	−0.012	0.015	0.012	0.013	0.181	0.148	0.153
$\alpha = -0.25$	0.010	0.011	0.005	0.007	0.005	0.006	0.083	0.066	0.078

$$h_i(t|\eta_{it}) = h_0(t)\exp\{\gamma_1 x_{i1} + \gamma_2 x_{i2} + \alpha\eta_{it}\} \quad \text{and} \quad h_0(t) = \lambda. \tag{5.2}$$

The parameters used in data generation are summarized in Table 1. The payment times are assumed to be exogenous and are set at $t = 0, 1, 2, ..., 9$. We assume $x_1 \sim Bernouli(0.5)$, representing a discrete predictor and $x_2 \sim Normal(1, 0.25)$, 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), we employ the Algorithm provided in the Online Appendix C to construct the training and validation data in the simulation study.

5.2. Parameter estimates

The main results on parameter estimation are summarized in Table 1. We 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 described in Section 4. The results reported in Table 1 are based on $S = 150$ replications.

We show in the table the average bias (Bias) and the average standard error (SE) of the estimates. In addition, we calculate the nominal standard deviation of the point estimates (SD) and report the standard deviation of the average bias calculated as $SD/\sqrt{(S)}$. As anticipated, both estimate and uncertainty of the average bias decrease as sample size increase. The average standard error are comparable to the nominal standard deviation, indicating the accuracy of variance estimates. Lastly, the standard errors are consistent with $\sqrt{n}$ convergence.

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

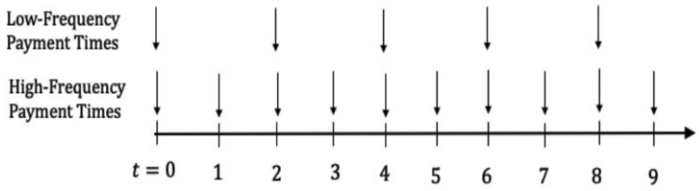

FIGURE 2: Payment times for low-frequency and high-frequency payment models.

To emphasize the importance of joint estimation, we 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. 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. Both estimation techniques turn out to introduce significant bias in the parameter estimates. Detailed discussions on the two alternative strategies and the associated estimation results are provided in the Online Appendix A.

5.3. 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 the independent and two-stage estimates. Results presented in this section are based on sample size $N = 1000$ and $S = 150$ replications.

In this scenario, we 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 5.1 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 2 illustrates the timeline of the payment times for the low-frequency and high-frequency payment models. It is seen that claims in the high-frequency model are likely to have more payment transactions than those in the low-frequency model. 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 model. However, a claim with the same settlement time will be closed with two transactions under the high-frequency payment model.

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

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

TABLE 2.

RBNS PREDICTION RESULTS UNDER HIGH AND LOW FREQUENCY PAYMENTS.

	High-frequency			Low-frequency		
$N = 1000, S = 150$	Mean	Error %	SE/$\sqrt{S}$	Mean	Error %	SE/$\sqrt{S}$
True reserve	6062		71	4412		49
Joint model error	33	0.55	74	51	1.16	55
Two-stage error	206	3.39	77	208	4.72	59
Independent error	−1583	−26.12	57	−908	−20.58	45

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 2 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 joint model 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 joint model 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, the joint model 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.

6. EMPIRICAL ANALYSIS USING THE JOINT MODEL

6.1. Data

The data analyzed in this paper are from the Wisconsin Local Government Property Insurance Fund (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,

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

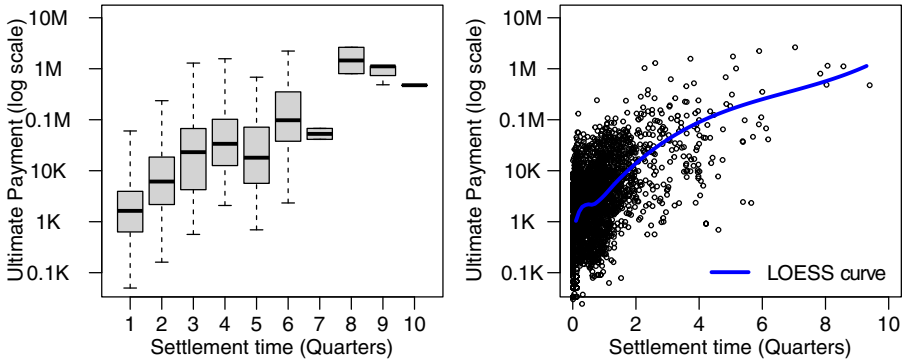

FIGURE 3: Relationship between settlement time and ultimate payment. The left panel shows the distribution of ultimate payment by settlement time. The right panel shows the scatter plot with fitted LOESS curve.

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.

We use data from building and coverage from January 1, 2006, to December 31, 2013, in the empirical analysis. For the purpose of reserve prediction, we assume a valuation date of December 31, 2009, which naturally split the claims data into two pieces. The training data contain claims that have occurred and been reported between January 1, 2006, and December 31, 2009. There are 3393 reported claims, among which, 129 open claims have no payments by the valuation date, and 34 claims with partial payments remain open by the valuation date. The validation data contain claims that are reported between January 1, 2006, and December 31, 2009, but settled between January 1, 2010, and December 31, 2013. Specifically, there are 163 claims with a total outstanding payment of $4,511,490. The training data are used to develop the joint model, and the validation data are used to evaluate the quality of reserve prediction. We emphasize that our analysis is based on ground-up losses, which allowed us to identify and exclude claims with reported losses less than the deductible, resulting in such claims being closed without payment. Therefore, our 3393 reported claims do not include claims which closed without payments.

Figure 3 visualizes the relationship between settlement time (in quarters) and the ultimate payments (in log scale) for claims in the training data set. The settlement time of a claim is defined as the difference between the close date and reporting date of the claim. For reopened claims, the settlement time of a claim is defined as the difference between the final close date after reopening and reporting date of the claim. The left panel shows the distribution of ultimate payment by settlement time, and the right panel shows the scatter plot of the two outcomes. The solid curve in the right panel corresponds to the fit of the

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

TABLE 3.

DESCRIPTION OF PREDICTORS IN THE JOINT MODEL.

Variable	Description
Claim/Transaction level covariates	
LnInitialEst	Initial case estimate in log scale
ReportDelay	Reporting time from occurrence in days
LossYear	Year of claim occurrence
LossQtr	Quarter of claim occurrence
CauseCode	A categorical variable to indicate the cause of claim
TimeToPayment	Payment time from reporting in days
Policy/Policyholder level covariates	
EntityType	A categorical variable to indicate the entity type: Village, City, County, School, Town or Miscellaneous
CountyCode	A categorical variable to indicate county of the entity
Region	A categorical variable to indicate the region: Northern, Northeastern, Southeastern, Southern, or Western
LnPolicyDed	Per-occurrence deductible in log scale

TABLE 4.

DESCRIPTIVE STATISTICS OF OUTCOMES AND PREDICTORS BASED ON CLOSED CLAIMS.

	Min.	Median	Mean	Max.	Ultimate loss (ρ_S)	Settlement time (ρ_S)
Ultimate loss	25	2203	14,133	2,633,822	–	0.49
Settlement time (days)	1	38	66	861	0.49	–
Deductible	500	1000	12,297	100,000	−0.28	−0.21
Initial estimate	30	2500	9545	1,000,000	0.93	0.51
Reporting delay (days)	0	28	66	864	−0.29	−0.55

locally estimated scatter plot smoother (LOESS). Both plots suggest a strong positive relation between ultimate payment and settlement time, that is, it takes longer to close larger claims. In addition, the data set contains rich information regarding the policy, policyholder, claims, and payment transactions. Table 3 describes the variables that we select to use as predictors in the joint model.

Table 4 provides descriptive statistics of the two outcomes of interest, that is, the settlement time and ultimate loss amount, as well as the continuous predictors. We note that the deductible and initial case estimate are right-skewed. To handle the skewness, we apply logarithmic transformation prior to model fitting. In addition, the table reports the Spearman correlation between the two outcomes, and then between each outcome and the predictors. The results suggest a substantial correlation between the settlement time and ultimate payment, and the selected covariates are expected to be predictive in the reserving application.

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

TABLE 5.

ESTIMATION RESULTS FOR THE JOINT MODEL.

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		κ	1.459	

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)

6.2. Estimation results

The joint longitudinal-survival framework is applied to the micro-level reserving problem using the property data from the Wisconsin LGPIF. Specifically, we use a gamma distribution with a log link and a nonlinear payment trend using B-splines with an internal knot at payment time 5 in the longitudinal submodel, and a Weibull baseline hazard in the survival submodel. The nonlinear payment trend is motivated by Figure 3, which suggests a nonlinear relationship between payment size and time. Also, a random intercept longitudinal submodel is assumed to follow a normal distribution, $\mathcal{N}(0, \nu)$. With the random intercept longitudinal submodel, only the intercept parameter in the GLMM is random, and all slope parameters are fixed.

The estimation results are presented in Table 5. We present the parameter estimates and standard errors for the continuous covariates. Here, $B_1 \ldots B_4$ denotes the spline coefficients in the longitudinal submodel. For the categorical covariates, we present the likelihood ratio test statistics, the degrees of freedom, and the associated p-values that indicate their statistical significance in each submodel.

In the survival submodel, the association parameter α is interpreted as the percentage change in hazard or risk of the settlement when expected cumulative payments increase by one percent. The estimated value is −0.407 and is

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

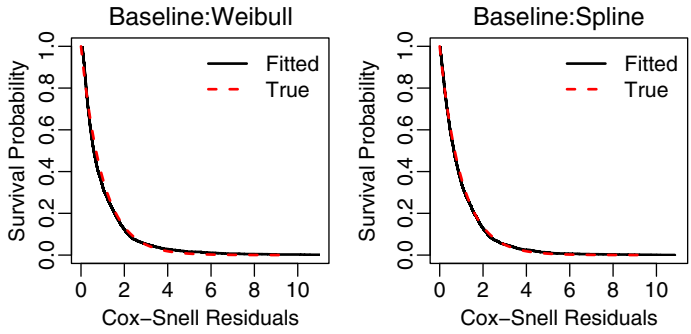

FIGURE 4: Visualization of goodness-of-fit of the survival submodel.

statistically significant at a 1% significance level. The negative association in the hazard model implies a positive relationship between the settlement time and payment amount.

6.2.1. Evaluation of survival submodel

The correct specification of the survival submodel is necessary to obtain accurate prediction results. In the modeling building process, we consider two alternative specifications for the baseline hazard function. A parametric Weibull model and a more flexible spline model. The spline baseline model was fitted with equally spaced five internal knots in the quantiles of the observed event times.

To examine the overall goodness-of-fit, we compare the Kaplan–Meier estimate of the Cox-Snell residuals from both survival submodels to the function of the unit exponential distribution (Rizopoulos, 2012). Figure 4 visualizes the fit for the survival submodel with the Weibull and spline baseline hazard functions in the left and right panel, respectively. 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 functions fit the data very well. We choose the Weibull model due to easy interpretation.

6.2.2. Evaluation of longitudinal submodel

We also explore alternative specifications in the longitudinal submodel. First, we investigate the distributional assumption for the payment amount. To accommodate the skewness in the claim severity, we fit two joint models, one with a lognormal and the other with a gamma longitudinal submodel. The two methods amount to the transformation technique and generalized linear models for handing non-normal responses in regression respectively. In the former, we transform the response to a symmetric distributed outcome and apply linear regression model to the transformed variable. In the latter, we use a link

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

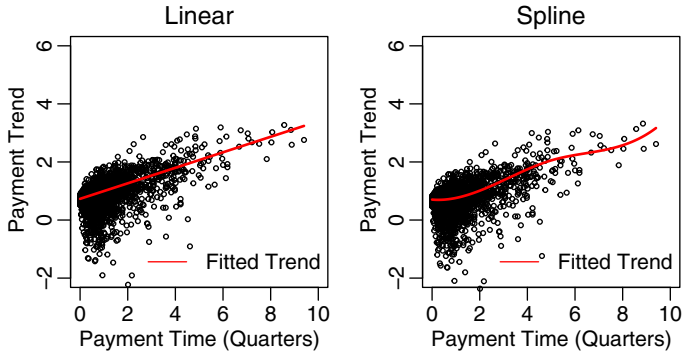

FIGURE 5: Evaluation of payment trend in the longitudinal submodel.

function to connect the mean of exponential family and the systematic component of covariates. See Shi (2014) for more detailed discussion. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for the joint model with the lognormal distribution are 74,117 and 74,488, respectively, and that of the gamma model are 73,887 and 74,258, respectively. The goodness-of-fit statistics suggest a better fit for the gamma model.

Second, we investigate the payment trend in the longitudinal submodel. To be more specific, we consider a linear trend and a nonlinear trend using splines. In Figure 5, we overlay the scatter plot of payments by time with the fitted trend. The left panel shows the linear trend, and the right panel shows the nonlinear trend using B-spline basis functions. The nonlinear trend shows a better fit, which is further supported by the AIC and BIC statistics that are not reported. As a result, we employed a gamma regression model with a nonlinear trend in the longitudinal submodel.

6.3. Reserve prediction

This section examines the reserve prediction from the proposed joint model. To recap, the validation data span from January 1, 2010 to December 31, 2013. All payments that occurred during this time period are the outstanding liabilities of the insurer as of the valuation date. Our goal in the reserving application is to predict the outstanding payments. One advantage of individual loss reserving models over macro-reserving methods is that we will be able to obtain not only the prediction for the insurance portfolio but also the prediction for each individual claim.

6.3.1. Prediction from the joint model

To obtain the RBNS reserve estimate from the fitted joint model, we follow the prediction routine described in Section 4.2. Specifically, the joint model allows

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

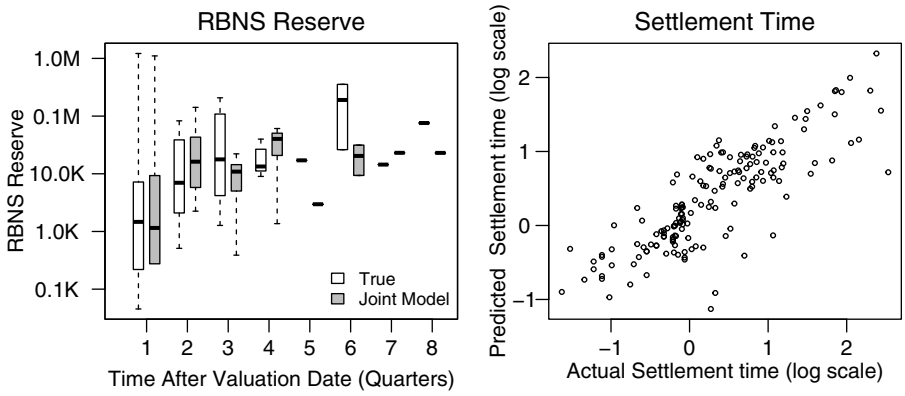

FIGURE 6: Comparison between actual and predicted values of unpaid payment and settlement time for individual claims.

us to make a prediction for both the time-to-settlement and the ultimate payment for individual claims. Given that the B-splines is used in the longitudinal submodel, prediction for the ultimate losses is based on a linear extrapolation for the settlement time beyond the maximum observed payment time in the training data.

Figure 6 compares the actual (or realized) and predicted outcomes in the hold-out sample. The left panel presents the distribution of unpaid losses over time. The consistency between the actual and predicted unpaid losses suggests a satisfying performance of the joint model. 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 shows the scatter plot of actual and predicted settlement time. The linear relationship is an indicator of prediction accuracy.

For reserving purposes, one is not only interested in the point prediction but also the predictive distribution. We discuss two predictive distributions for the insurer's outstanding payments that are relevant to the application. The first is the predictive distribution of the expected outstanding payments. The uncertainty associated with the expected payments is from the parameter estimation. This type of uncertainty is known as parameter uncertainty in the macro-loss reserving literature (see, for instance, England and Verrall, 2002). The second is the predictive distribution of the outstanding payments. In addition to parameter estimation, additional uncertainty arises due to the inherent randomness of the unpaid losses, which is known as process uncertainty in the macro-loss reserving literature.

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

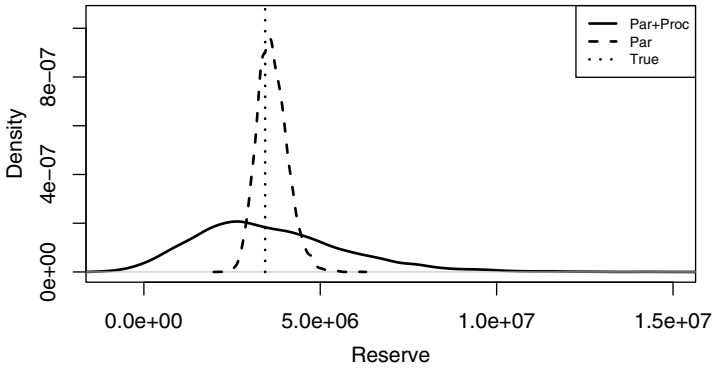

FIGURE 7: Predictive distributions of the total RBNS reserve.

We obtain the two types of predictive distribution using simulation to incorporate the parameter and process uncertainty. For parameter uncertainty, we generate model parameters from the multivariate normal distribution with mean being the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ and covariance matrix $\widehat{Var}(\hat{\boldsymbol{\theta}})$. For process uncertainty, we generate the ultimate payment from the joint model given the simulated parameters. To illustrate, we present in Figure 7 the predictive distributions for the entire insurance portfolio. The RBNS liability is generated for each individual claim and then aggregated for the portfolio. As anticipated, the predictive distribution of the outstanding payment is much wider than the predictive distribution of the expected payment because the former contains both parameter and process uncertainty, while the latter only considers parameter uncertainty. The vertical line in the figure indicates the actual outstanding payments in the hold-out sample. It corresponds to the 34th percentile and 53rd percentile in the predictive distributions of expected unpaid loss and unpaid loss, respectively.

6.3.2. Comparison with alternative methods

This section compares reserve prediction from the joint model with alternative individual loss reserving methods. Specifically, we consider the four methods below:

1. Independence Model: This model assumes $\alpha = 0$ in the proposed joint model. Hence, it is nested by the joint model assuming independence between the longitudinal and survival submodels.
2. Two-stage method: This method differs from the proposed joint model in estimation method. Specifically, the first stage estimates the parameters in the longitudinal submodel, and the second stage estimates the parameters in the survival submodel holding parameter estimates from the first stage fixed.

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

3. GLM: This model contains two components. One component uses a gamma GLM for ultimate payments of a claim, and the other components employs a survival model for the settlement time of the claim. The two components are assumed to be independent and are estimated separately. The model only consider closed claims.
4. MPP: This framework considers the entire claim process, including occurrence, reporting, and development after reporting. A counting process is used to model the occurrence of claims. Upon occurrence, the transaction time, the type of transaction, and the transaction's payment amount are considered to be the marks (features of interest). 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. We follow the prediction routine for the RBNS reserve in Antonio and Plat (2014). The prediction routine simulates the transaction time, type of the transaction (payment to a settlement, or intermediate payment), and the corresponding payment amount. Detailed model specifications for the MPP are provided in the Online Appendix B.

In addition, we make a comparison with the reserves determined by a claim expert (initial estimate). Further, we provide a comparison of reserves on the aggregate level and provide reserve estimates from the chain-ladder model. We employ the Mack chain-ladder model (Mack, 1993) and obtain RBNS reserve estimates from a run-off triangle that is aggregated using the reporting and observation year instead of the occurrence year and development year. The analysis was performed in the ChainLadder R package (Carrato *et al*., 2020).

The comparison is based on predictions of unpaid losses for individual claims. Using the actual and predicted unpaid losses of individual claims, we calculate five validation statistics, mean absolute error (MAE), root mean squared error (RMSE), Pearson correlation, Gini correlation, and simple Gini (see Frees *et al*., 2011 and Frees *et al*., 2014 for details on the latter two metrics). The results are summarized in Table 6. Higher prediction accuracy is suggested by a smaller MAE and RMSE, and larger correlations. The validation statistics support the superior prediction from the proposed joint model to alternative individual reserving methods. We also calculate the reserve error on the aggregate level, which is the expected RBNS reserve minus the actual unpaid losses, as a percentage of the actual unpaid losses. The results show the chain-ladder method did not perform well in estimating the unpaid losses.

6.4. Limitations

Though the joint model displayed superior prediction results compared to models that ignore the payment–settlement association; there are some limitations in the current framework, which will be the subject of future research and are highlighted below:

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

TABLE 6.

COMPARISON OF PREDICTIONS OF UNPAID LOSSES FOR INDIVIDUAL CLAIMS.

	MAE	RMSE	Pearson r (%)	Gini Cor (%)	Gini	Total RBNS Reserve error (%)
Joint model	24,396	77,419	74.86	29.24	18,467	−1.37
MPP	30,296	115,954	34.04	7.61	4,805	1.28
Two-stage	42,792	185,995	23.59	23.01	14,531	35.53
Independent (JM with $\alpha = 0$)	37,959	169,556	−14.05	6.96	4,397	−3.26
GLM (closed claims)	76,230	355,144	43.31	25.54	16,131	224.81
Initial estimate (by claim expert)	22,260	97,173	47.24	29.56	18,643	−32.41
Chain-Ladder	–	–	–	–	–	27.97

1. Working with cumulative instead of incremental payments: From equation (3.3), we explicitly capture the dependence between the payment history and settlement process. To do this, we rely on the assumption that the cumulative payments follow a continuous growth curve. The implicit assumption is that the trends in time, fixed-effects, and the random effects $\boldsymbol{b}_i$ together will soak up the temporal correlation among cumulative payments. The conditional independence assumption for cumulative payments could be too strong to result in a realistic claim evolution. One could explicitly account for the correlation in cumulative payments but at the cost of increased model complexity. We leave this work to explore in the future.
2. Intermediate payment prediction: The model immediately predicts the amount paid at settlement without intermediate cash flows, which may be limiting in situations where such quantities are needed.
3. Time-dependent covariates: The proposed model only allows for external time-varying covariates that are determined independently of the settlement process. Internal time-varying covariates such as case estimates, whose value is generated until settlement or censoring by the individual claims, require special treatment (Kalbfleisch and Prentice, 2002). Handling internal time-varying covariates is a topic for future research.

7. CONCLUSION

This paper concerns the claims reserving problem using an individual-level loss reserving method. The work was primarily motivated by the payment–settlement association. Specifically, complex claims can be both

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

more expensive in terms of severity 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.

We introduced a joint model framework to the individual-level loss reserving literature to accommodate such association. The joint model consists of a longitudinal submodel for the cumulative payment process and a survival submodel for the settlement process, and the association between the two components is induced via a shared parameter model. In addition, the proposed 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.

We have demonstrated that failing to incorporate the association between the payment processes and the settlement processes could lead to significant errors in reserving prediction. Specifically, 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 were compared to existing reserving models, and the results showed that accounting for the payment–settlement association leads to better prediction and lower reserve uncertainty compared to models that ignore the association.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2021.28

References

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.

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.

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.

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.

ELASHOFF, R.M., LI, G. and LI, N. (2017) *Joint Modeling of Longitudinal and Time-to-Event Data*. Boca Raton, FL: Chapman and Hall/CRC.

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., MEYERS, G. and CUMMINGS, A.D. (2014) Insurance ratemaking and a gini index. *Journal of Risk and Insurance* **81**(2), 335–366.
FREES, E.W., MEYERS, G. and CUMMINGS, D. (2011) Summarizing insurance scores using a gini index. *Journal of the American Statistical Association*, **106**(495), 1085–1098.
FRIEDLAND, J.F. (2010) *Estimating Unpaid Claims Using Basic Techniques*. Casualty Actuarial Society. http://www.casact.org/sites/default/files/database/studynotes_friedland_estimating.pdf.
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. In Casualty Actuarial Society Forum, pp. 139–210.
KALBFLEISCH, J. and PRENTICE, R. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edn. NewYork: Wiley.
LANGE, K. (2004) *Optimization*. New York: Springer-Verlag.
LITTLE, R. (2008) Selection and pattern-mixture models. In *Longitudinal Data Analysis* (eds. G. FITZMAURICE, M. DAVIDIAN, G. VERBEKE and G. MOLENBERGHS), chapter 18, pp. 409–432. Boca Raton: CRC Press.
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.
LOPEZ, O., MILHAUD, X. and THÉROND, P.-E. (2019) A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin*, **49**(3), 741–762.
MACK, T. (1993) Distribution free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, **23**(2), 213–225.
MOLENBERGHS, G. and VERBEKE, G. (2006) *Models for Discrete Longitudinal Data In: Springer Series in Statistics*. New York: Springer.
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.
PETRONI, K.R. (1992) Optimistic reporting in the property-casualty insurance industry. *Journal of Accounting and Economics*, **15**(4), 485–508.
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*. Boca Raton, FL: 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. (2014) Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science. Volume 1: Predictive Modeling Techniques* (eds. E.W. FREES, R.A. DERRIG and G. MEYERS), pp. 236–259. Cambridge, MA: Cambridge University Press.
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.

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press

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. 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.
TSIATIS, A. and DAVIDIAN, M. (2004) Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14**(3), 809–834.
VERBEKE, G., MOLENBERGHS, G. and RIZOPOULOS, D. (2010) Random effects models for longitudinal data. In *Longitudinal Research with Latent Variables* (eds. K. VAN MONTFORT, J. H. OUD and A. SATORRA), chapter 2, pp. 37–96. Berlin, Heidelberg: Springer.
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*. Chichester, West Sussex: 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.

A. NII-ARMAH OKINE (Corresponding author)
Department of Mathematical Sciences
Appalachian State University
121 Bodenheimer Dr, Boone, NC 28608, USA
E-mail: okinean@appstate.edu

EDWARD W. FREES
Department of Risk and Insurance
Wisconsin School of Business, University of Wisconsin-Madison
975 University Avenue, Madison WI 53706, USA
E-Mail: jfrees@bus.wisc.edu

PENG SHI
Department of Risk and Insurance
Wisconsin School of Business, University of Wisconsin-Madison
975 University Avenue, Madison WI 53706, USA
E-Mail: pshi@bus.wisc.edu

https://doi.org/10.1017/asb.2021.28 Published online by Cambridge University Press