Lognormal Mixed Models for Reported Claim Reserves

Katrien Antonio *† Jan Beirlant ‡ Tom Hoedemakers § Robert Verlaak ¶

Revision August 2005

Abstract

Traditional claims reserving techniques are based on so-called run-off triangles containing aggregate claim figures. Such a triangle provides a summary of an underlying data set with individual claim figures. In this contribution the authors explore the interpretation of the available individual data in the framework of longitudinal data analysis. Making use of the theory of linear mixed models, a flexible model for loss reserving is built. Whereas traditional claims reserving techniques don't lead directly to predictions for individual claims, the mixed model enables such predictions on a sound statistical basis. Both a likelihood-based as well as a Bayesian approach are considered. In the frequentist approach, expressions for the mean squared error of prediction of an individual claim reserve, origin year reserves and the total reserve are derived. Using MCMC techniques, the Bayesian approach allows to simulate the complete predictive distribution of the reserves and the calculation of various risk measures derived from it. The article ends with an illustration of the suggested techniques on a data set from practice. The results are compared with those obtained with traditional claims reserving techniques for run-off triangles.

Keywords: loss reserving, mixed models, longitudinal data, prediction, Bayesian statistics.

*Corresponding author: Katrien.Antonio@econ.kuleuven.be

†Ph. D. student, University Center for Statistics, KU Leuven, W. de Croylaan 54, 3001 Heverlee

‡University Center for Statistics, KU Leuven, W. de Croylaan 54, 3001 Heverlee

§Department of Applied Economics, Naamsestraat 69, 3000 Leuven

¶AON Re Belgium, Van Nieuwenhuyselaan 2, 1160 Brussel

1 Introduction

Claims originating in a particular year often can not be finalized in the same year. Many causes for delay of the payment process are possible, for example long-lasting juridical procedures are the rule with liability insurance. Alongside the *reported but not settled* (RBNS) claims, a company also needs to manage claims that *incurred* already *but* are *not* yet *reported* (IBNR) to the insurer. For both types of claims, provisions will be held to meet the future obligations of the insurer towards its policy holders. In this contribution we concentrate on the prediction of remaining payments for reported claims.

A broad literature is available concerning deterministic and stochastic models used for loss reserving. We refer to England & Verrall (2002) for an overview. The methods discussed by these authors are framed within the context of a run-off triangle like the one in Table 1. The random variable Y_{ij} (for $i, j = 1, \ldots, t$) denotes the claim figure for year of origin (arrival or incurral year) i and development year j, made up by aggregating the individual claims corresponding with this (i, j) combination. For (i, j) cells with $i + j \leq t + 1$, Y_{ij} has already been observed, otherwise it is a future observation. As well as incremental, cumulative or incurred payments, these random variables can denote quantities such as loss ratios. The purpose of loss reserving techniques is to complete this run-off triangle to a square and even to a rectangle if estimates are required pertaining to development years for which no data are recorded in the run-off triangle at hand.

Arrival	*Development Year*						
Year	1	2	$\ldots$	j	$\ldots$	$t-1$	t
1	Y_{11}	Y_{12}	$\ldots$	Y_{1j}	$\ldots$	$Y_{1,t-1}$	Y_{1t}
2	Y_{21}	Y_{22}	$\ldots$	Y_{2j}	$\ldots$	$Y_{2,t-1}$	
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$		
i	Y_{i1}	$\ldots$	$\ldots$	Y_{ij}			
$\vdots$	$\ldots$	$\ldots$	$\ldots$				
t	Y_{t1}						

Table 1: *Random variables in a run-off triangle.*

The present literature on loss reserving only contains techniques based on summary triangles like the one in Table 1. However, some authors recently suggested to leave the track of aggregate claim figures. To illustrate this statement we quote England & Verrall (2002, p.507): "*The problem is more with the data than the methods, since, clearly, it is the estimation of aggregate case reserves which is at fault. [...] In this respect, models based on individual claims, rather than data aggregated into triangles, are likely to be of benefit.*" In some recent publications Taylor *et al.* (2002, 2003) put forward this same idea as the future of loss reserving techniques: "*The triangle is a summary, whose origins are* *very much driven by the computational restrictions of a bygone era. [...] Indeed, one can imagine future generations of students, educated on the basis of such models, finding the compression of data into a triangle quite artificial.*" (Taylor *et al.*, 2002, p.21). Inspired by these quotations, the intention of this contribution is to present a statistical framework to model data sets containing individual records. Based on the work of Norberg (1993), Haastrup & Arjas (1996) suggested to model the data from individual claims in a non-parametric Bayesian way with the occurrence and development of claims modelled as marked point processes. Our contribution interprets the data from individual claims as longitudinal data and uses the concept of general linear mixed models as a tool to model them, both in a likelihood-based and a Bayesian way.

Focusing on the complete individual record data underlying a run-off triangle, the analyst is assumed to have a data set at hand with the following characteristics (see also Taylor *et al.*, 2003)

(i) a unique reference to denote each claim in the data set;

(ii) a record for each payment made for a particular claim and (if available) each change in the company's estimate of the incurred loss;

(iii) the arrival and reporting year for each claim, the development and calendar year to which a payment belongs;

(iv) (if available) information concerning specific features of the policyholder (age, gender, etc.).

To represent such an extensive individual data set, let the random variable $Y(i,k,j)$ denote the claim figure for the k^{th} claim from arrival year i in its j^{th} development year. The number of claims in arrival year i is denoted by n_i and $t(i,k)$ denotes the development year of the last observation for the k^{th} claim from arrival year i. As in Table 1, the figures represented by the random variables can be for instance incremental, cumulative or incurred payments or loss ratios.

For every claim in a unit record data set, repeated measurements (e.g. on incremental, cumulative or incurred payments) are taken over a certain period of time, namely the development of the claim. In this way it appears natural to interpret the available data in the context of longitudinal data analysis, i.e. the analysis of repeated measurements on a group of subjects over time. This stands in contrast to cross-sectional data where a response is measured only once per subject. One class of models for longitudinal data are linear mixed models. Mixed models for longitudinal data became very popular after the appearance of the paper by Laird & Ware (1982). In this paper we explain how individual record data sets can be modelled in the context of mixed models and how these models lead to forecasts for future payments for the reported but not completely settled claims. Important to point out is that the logarithm of the individual data is modelled and not the data on the original scale. In this way our models are to be compared with the well-known lognormal regression models for loss reserving (as summarized in Section 3). Applications of mixed models in the context of credibility theory have been described before by Frees *et al.* (1999, 2001). This text presents another actuarial domain, namely loss reserving, where the models can be useful.

The problem of individual reserving is often encountered in the reinsurance business. The authors' experience learns that practitioners most of the time apply classical chain-ladder development factors (based on an aggregate analysis) to forecast future payments of individual claims. This contribution presents an alternative approach which is based on a sound statistical analysis of the data. Section 4 illustrates the presented technique on a data set from practice and compares our results with those obtained using classical claims reserving techniques, both on the aggregate and individual level.

The rest of the paper has been organized as follows. Section 2 motivates the use of the general linear mixed model and contains the necessary statistical background. Section 3 resumes some well-known lognormal regression models that are widely used in claims reserving. Afterwards it is described how loss reserving based on individual record data sets can be performed within the framework of linear mixed models, both in a likelihood and a Bayesian way. The paper ends with the illustration of the presented techniques on the data set from practice.

2 General linear mixed models

This section motivates the use of general linear mixed models to analyze an individual record data set as described in Section 1. Alongside this, an introduction to the concepts of mixed models is given. For more statistical details we refer to Verbeke & Molenberghs (1997, 2000) or Demidenko (2004).

2.1 Motivation

The interpretation of observations on individual claims as longitudinal data was already motivated in Section 1. Obviously, models for the longitudinal data from an individual record data set have to fulfill certain requirements. Firstly, it should be clear that the number of observed payments per claim is not necessarily the same for every claim in the unit record data set. Different claims are also observed at different stages in their development. In the context of longitudinal data one speaks about 'unbalanced data'. Therefore, statistical models are needed which allow the number of measurements (here: payments or loss ratios) and times of observations to vary among subjects (here: claims). Secondly, methods are needed that enable to model the dependencies among the observations on a certain subject/claim. Imagine e.g. that individual cumulative payments are modelled, then observations on the same claim can not be assumed to be independent. General linear mixed models are one class of models for longitudinal data that fulfill these requirements. Moreover, when using mixed models, the deviation of a particular payment profile from the global average can be modelled explicitly by the inclusion of claim-specific random effects in the model specification.

2.2 Statistical background

Linear mixed models extend classical linear models by incorporating random effects in the structure for the mean. Assume that the data set at hand consists of N subjects (here –again– claims). Let n_i denote the number of observations for the i^{th} subject. $\boldsymbol{Y}_i$ is the $n_i \times 1$ vector of observations for the i^{th} claim ($1 \leq i \leq N$). The general linear mixed model is given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \tag{1}$$

$\boldsymbol{\beta}$ ($p \times 1$) gives the p fixed effects parameters. These are fixed, but unknown, regression parameters, common to all subjects. $\boldsymbol{b}_i$ ($q \times 1$) is the vector with the random effects parameters for the i^{th} subject in the data set. The use of random effects reflects the belief that there is heterogeneity among subjects for a subset of the regression coefficients in $\boldsymbol{\beta}$. $\boldsymbol{X}_i$ ($n_i \times p$) and $\boldsymbol{Z}_i$ ($n_i \times q$) are the design matrices for the p fixed and q random effects. $\boldsymbol{\epsilon}_i$ ($n_i \times 1$) contains the residual components for subject i. Independence between subjects is assumed. $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are also assumed to be independent and we follow the traditional assumption that they are normally distributed with mean vector 0 and covariance matrices, say $\boldsymbol{D}$ ($q \times q$) and $\boldsymbol{\Sigma}_i$ ($n_i \times n_i$), respectively. Different structures for these covariance matrices are possible; an overview of some frequently used ones can be found in e.g. Verbeke & Molenberghs (1997, 2000). It is easy to see that $\boldsymbol{Y}_i$ then has a marginal normal distribution with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{V}_i = \text{Var}(\boldsymbol{Y}_i)$, given by

$$\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{'} + \Sigma_i. \tag{2}$$

In this interpretation it becomes clear that the fixed effects enter only the mean $\text{E}[\boldsymbol{Y}_i]$, whereas the inclusion of subject-specific effects specifies the structure of the covariance between observations on the same unit or claim. Under the traditional normality assumptions,

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{b}_i &\sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i) \\ \boldsymbol{b}_i &\sim N(0, \boldsymbol{D}), \end{aligned} \tag{3}$$

it becomes clear that the residual terms model variability within a subject.

Denote the unknown parameters in the covariance matrix $\boldsymbol{V}_i$ with $\boldsymbol{\alpha}$. Conditional on $\boldsymbol{\alpha}$, a closed form expression for the maximum likelihood estimator of $\boldsymbol{\beta}$ exists, namely

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{Y}_i. \quad (4)$$

Conditional on $\boldsymbol{\alpha}$, this is the Best Linear Unbiased Estimator (BLUE) for $\boldsymbol{\beta}$, where 'best' is in the sense of minimum mean squared error. To predict the random effects, the mean of the posterior distribution of the random effects given the data, $\boldsymbol{b}_i|\boldsymbol{Y}_i$, is used. Conditional on $\boldsymbol{\alpha}$, we have

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}), \quad (5)$$

which can be proven to be the Best Linear Unbiased Predictor (BLUP) of $\boldsymbol{b}_i$ (where 'best' is again in the sense of minimum mean squared error). Estimation of $\boldsymbol{\alpha}$ is mostly performed by maximum likelihood (ML) or restricted maximum likelihood (REML). The expression maximized by the ML (L_1), respectively REML (L_2), estimates is given by

$$L_1(\boldsymbol{\alpha}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = c_1 - \frac{1}{2} \sum_{i=1}^{N} \log |\boldsymbol{V}_i| - \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{r}_i' \boldsymbol{V}_i^{-1} \boldsymbol{r}_i \quad (6)$$

$$L_2(\boldsymbol{\alpha}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = c_2 - \frac{1}{2} \sum_{i=1}^{N} \log |\boldsymbol{V}_i| - \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{r}_i' \boldsymbol{V}_i^{-1} \boldsymbol{r}_i - \frac{1}{2} \sum_{i=1}^{N} \log |\boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i|, \quad (7)$$

where $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{y}_i\right)$ and c_1, c_2 are appropriate constants. (6) and (7) are maximized using iterative numerical techniques such as Fisher scoring or Newton-Raphson (for full details, see e.g. Demidenko, 2004). In (4) and (5) the unknown $\boldsymbol{\alpha}$ is then replaced with $\hat{\boldsymbol{\alpha}}_{ML}$ or $\hat{\boldsymbol{\alpha}}_{REML}$, leading to the empirical BLUE for $\boldsymbol{\beta}$ and the empirical BLUP for $\boldsymbol{b}_i$. For inference regarding the fixed and random effects and the variance components, appropriate likelihood ratio and Wald tests are explained in Verbeke & Molenberghs (2000).

The predicted value for the conditional expectation $\boldsymbol{Y}_i^{\star} := \mathrm{E}[\boldsymbol{Y}_i|\boldsymbol{b}_i] = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ is obtained from (4) and (5), namely

$$\begin{aligned}
\hat{\boldsymbol{Y}}_i^{\star} &= \boldsymbol{X}_i \widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i \widehat{\boldsymbol{b}}_i \\
&= \boldsymbol{X}_i \widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' \boldsymbol{V}_i^{-1}) \boldsymbol{X}_i \widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' \boldsymbol{V}_i^{-1} \boldsymbol{Y}_i \\
&= \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \widehat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i} - \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1}) \boldsymbol{Y}_i.
\end{aligned}$$

Note that this expression can be interpreted as a *credibility* predictor, because it is a weighted average of $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ (related to the whole database) and $\boldsymbol{Y}_i$ (related to subject i). The credibility weights are $\boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1}$ and $\boldsymbol{I}_{n_i} - \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1}$, which implies that $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ gets much weight if the residual variability is 'large' in comparison with the total variability. $\widehat{\boldsymbol{Y}}_i^{\star}$ is the BLUP of $\boldsymbol{Y}_i^{\star}$. If the residual terms are modelled independently (thus $\boldsymbol{\Sigma}_i$ diagonal for every i), $\widehat{\boldsymbol{Y}}_i^{\star}$ is also the BLUP for $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$.

Verbeke & Molenberghs (1997) describe in a detailed way how different types of mixed models can be fit with the statistical software package SAS [1]. For this paper we also used Proc Mixed from SAS to do the likelihood analysis in Section 4. For the Bayesian approach to mixed models, the previously mentioned distributional assumptions are used, together with a prior specification for the unknown parameters. A Gibbs sampling scheme is then set up to sample from the relevant posterior and predictive distributions. More details are given in Section 3.2. WinBUGS [2] is used for a specific analysis in Section 4. The availability of standard statistical software packages to analyze longitudinal data and fit mixed models, together with the diagnostic and graphical tools they provide, are important advantages that favour the use of mixed models in a practical loss reserving context.

3 Reported claim reserving using lognormal mixed models

The mixed models for individual loss reserving combine the ideas of general linear mixed models (see Section 2) with those of lognormal regression models for claims reserving (see below for a brief overview). The use of random effects allows to fit a claim-specific payment profile, by adding claim-specific behavior to the global payment pattern described by the fixed effects structure. When analyzing a concrete data set in Section 4, possible choices for the fixed and random effects are discussed.

As mentioned earlier, only mixed models for the logarithmic transformed individual data are considered. We briefly review the lognormal regression models that are widely used for loss reserving based on run-off triangles. Working on the logarithmic scale, claim figures must be strictly positive. Kunkler (2004) recently presented a possible approach to the (Bayesian) modelling of zero payments in a lognormal regression model for aggregate data. In further work, the mixed models presented here can be extended to other distributional frameworks and can be adopted to model zero or negative incremental payments or censored data by using two-parts models based on generalized linear mixed models (GLMM).

Applied to a run-off triangle like the one in Table 1, the general lognormal regression

[1]SAS is a commercial software package (for details see http://www.sas.com).

[2]WinBUGS is an open domain software package and is part of the Bayesian inference Using Gibbs Sampling project (for details see http://www.mrc-bsu.cam.ac.uk/bugs/).

model is given by

$$\log(\boldsymbol{Y}) = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}) \tag{8}$$

where $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1t}, \ldots, Y_{t1})'$ denotes the observed part of the run-off triangle. This implies that $\boldsymbol{Y}$ follows a lognormal distribution, namely $\boldsymbol{Y} \sim LN(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. The lognormal model with chain-ladder type structure for the mean (Kremer, 1982)

$$\log(Y_{ij}) = \alpha_i + \beta_j + \epsilon_{ij}, \tag{9}$$

is a first example of a widely known lognormal regression model for loss reserving. Hereby the α_i's are parameters for the arrival years and the β_j's for the development years. For a general model with parameters in the three directions (arrival, development and calendar year), we refer to De Vylder & Goovaerts (1979). Some special cases are the Probabilistic Trend Family (PTF) of models (Barnett and Zehnwirth, 1998), where

$$\log(Y_{ij}) = \alpha_i + \sum_{l=1}^{j-1} \beta_l + \sum_{t=1}^{i+j-2} \gamma_t + \epsilon_{ij}, \tag{10}$$

with the γ_t parameters for calendar year effects, and the Hoerl curve as in Zehnwirth (1985), with

$$\log(Y_{ij}) = \alpha_i + \beta \log(j) + \gamma j + \epsilon_{ij}. \tag{11}$$

To set up the reserves in a run-off triangle, one has to forecast the lower triangle in Table 1, namely the Y_{ij}'s with $i + j > t + 1$.

Recall the notation introduced in Section 1 to describe an individual claim data set. $Y(i,k,j)$ denotes what has been paid in or up to development year j for the k^{th} claim from arrival (or incurral) year i. Let n_{ik} denote the number of observations available for the k^{th} claim from arrival year i. $\boldsymbol{Y}(i,k) = \{Y(i,k,1), \ldots, Y(i,k,n_{ik})\}'$ is the $n_{ik} \times 1$ vector that contains the historical data for this claim. Our model for individual loss reserving is a lognormal mixed model, specified as

$$\begin{aligned} \boldsymbol{W}(i,k) := \log(\boldsymbol{Y}(i,k)) &= \boldsymbol{X}(i,k)\boldsymbol{\beta} + \boldsymbol{Z}(i,k)\boldsymbol{b}(i,k) + \boldsymbol{\epsilon}(i,k) \\ \boldsymbol{b}(i,k) &\sim N(0, \boldsymbol{D}) \\ \boldsymbol{\epsilon}(i,k) &\sim N(0, \boldsymbol{\Sigma}(i,k)). \end{aligned} \tag{12}$$

(i,k) refers to the k^{th} claim from arrival (or incurral) year i. $\boldsymbol{\beta}$ $(p \times 1)$ contains the fixed effects and $\boldsymbol{X}(i,k)$ $(n_{ik} \times p)$ is the corresponding design matrix. $\boldsymbol{b}(i,k)$ $(q \times 1)$ refers to the random effects and $\boldsymbol{Z}(i,k)$ $(n_{ik} \times q)$ to their design matrix.

The goal of the proposed statistical model is the prediction of outstanding payments for reported claims, on the basis of a data set with individual claim figures. Denote with $Y(i,k,u)$ a future payment for the k^{th} claim from arrival year i. Here, u, from *unobserved*, points at a development year $j > t(i,k)$. On the logarithmic scale, the random variables that need to be predicted, are given by

$$W(i,k,u) := \log\left(Y(i,k,u)\right) = \boldsymbol{x}(i,k,u)'\boldsymbol{\beta} + \boldsymbol{z}(i,k,u)'\boldsymbol{b}(i,k) + \epsilon(i,k,u). \tag{13}$$

In (13), $\boldsymbol{x}(i,k,u)$ and $\boldsymbol{z}(i,k,u)$ are the $p \times 1$ and $q \times 1$ covariate vectors for the fixed and random effects, respectively. Dealing with incremental payments the individual reserve for the claim under consideration is $\sum_u Y(i,k,u)$. For cumulative payments the individual reserve becomes $Y(i,k,T) - y(i,k,t(i,k))$, with T the time horizon in the direction of development years, $y(i,k,t(i,k))$ the last observed cumulative payment for this claim and $t(i,k)$ as in Section 1. Obviously, these expressions can be generalized to arrival year reserves or the total reserve for the portfolio.

In the sequel of this section both a likelihood-based and a Bayesian analysis of the suggested models for individual loss reserving are discussed. In the likelihood framework, expressions for estimates of the reserves on different levels (individual, year of origin and total), together with an estimate of their prediction error, are derived. A Bayesian analysis of the mixed claims reserving models allows simulation from the full predictive distribution of the different reserves and the empirical calculation of different risk measures. An illustration of the techniques is given in Section 4.

3.1 Likelihood-based approach: estimates of the reserves and prediction errors

To predict $W(i,k,u)$ in (13) (and afterwards $Y(i,k,u)$), the likelihood approach uses the BLUP for $W^\star(i,k,u) := \mathrm{E}[W(i,k,u)|\boldsymbol{b}(i,k)]$. From Section 2 we know that this is given by

$$\widehat{W}^\star(i,k,u) \quad = \quad \boldsymbol{x}(i,k,u)'\widehat{\boldsymbol{\beta}} + \boldsymbol{z}(i,k,u)'\widehat{\boldsymbol{b}}(i,k), \tag{14}$$

with $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}(i,k)$ similar to (4) and (5), but adjusted for the specific set-up of the extensive data set on individual claims. $\widehat{W}^\star(i,k,u)$ is an unbiased predictor for both $W^\star(i,k,u)$ and $W(i,k,u)$, in the sense that the expectations of these random variables are equal.

Following Frees *et al.* (1999) the Mean Squared Error of Prediction is given by:

$$\begin{aligned}
& \mathrm{E}[\widehat{W}^\star(i,k,u) - W^\star(i,k,u)]^2 = \mathrm{Var}[\widehat{W}^\star(i,k,u) - W^\star(i,k,u)] \\
= \; & \Big(\boldsymbol{x}(i,k,u)' - \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{X}(i,k)\Big) \\
\times \; & \left(\sum_h \boldsymbol{X}_h'\boldsymbol{V}_h^{-1}\boldsymbol{X}_h\right)^{-1} \times \Big(\boldsymbol{x}(i,k,u)' - \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{X}(i,k)\Big)' \\
- \; & \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{Z}(i,k)\boldsymbol{D}\boldsymbol{z}(i,k,u) + \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{z}(i,k,u),
\end{aligned} \tag{15}$$

where the index h runs over all claims in the data set. An analogous expression for the Mean Squared Error of Prediction $\text{E}[\hat{W}^\star(i,k,u) - W(i,k,u)]^2 = \text{Var}[\hat{W}^\star(i,k,u) - W(i,k,u)]$ is obtained by replacing $\boldsymbol{z}(i,k,u)'\boldsymbol{DZ}(i,k)'$ in (15) with $\boldsymbol{z}(i,k,u)'\boldsymbol{DZ}(i,k)' + \text{Cov}(\epsilon(i,k,u), \boldsymbol{\epsilon}(i,k))$ and $\boldsymbol{z}(i,k,u)'\boldsymbol{Dz}(i,k,u)$ with $\boldsymbol{z}(i,k,u)'\boldsymbol{Dz}(i,k,u) + \text{Var}(\epsilon(i,k,u))$. In case of independent residual terms (and thus all $\boldsymbol{\Sigma}(i,k)$ diagonal) this expression reduces to (15)+Var($\epsilon(i,k,u)$). Both expressions for the MSEP are conditional on the unknown variance components in $\boldsymbol{D}$ and $\boldsymbol{\Sigma}(i,k)$. In practice, these are estimated, e.g. with REML, and are plugged into the appropriate covariance matrices.

So far, only predictions on the logarithmic scale are considered. Predictions for individual profiles on the original scale of the payments are obtained by taking the characteristics of the lognormal distribution into account. The following expressions are used:

$$\begin{aligned}
\hat{Y}_{GeoM}(i,k,u) &= \exp\left\{\widehat{W}^\star(i,k,u)\right\}, \\
\hat{Y}_{Mean}(i,k,u) &= \exp\left\{\widehat{W}^\star(i,k,u) + \frac{1}{2}\text{Var}\left(W(i,k,u) - \widehat{W}^\star(i,k,u)\right)\right\}, \\
\widehat{\text{Var}}(Y(i,k,u)) &= \hat{Y}^2_{Mean}(i,k,u)\left[\exp\left\{\text{Var}(W(i,k,u) - \widehat{W}^\star(i,k,u))\right\} - 1\right], \quad (16)
\end{aligned}$$

where GeoM stands for geometric mean.

When dealing with incremental payments, the mean of an individual reserve $\sum_u Y(i,k,u)$ is estimated by

$$\sum_u \exp\left\{\widehat{W}^\star(i,k,u) + \frac{1}{2}\text{Var}(\widehat{W}^\star(i,k,u) - W(i,k,u))\right\}. \quad (17)$$

To get an idea of the variability of such an individual reserve an estimator is needed for $\text{Var}[\sum_u Y(i,k,u)]$. According to the characteristics of the lognormal distribution, the following expression is used:

$$\begin{aligned}
&\sum_u \hat{Y}^2_{Mean}(i,k,u)\left[\exp\left\{\text{Var}(W(i,k,u) - \widehat{W}^\star(i,k,u))\right\} - 1\right] \\
+ &\sum_u \sum_{u' \neq u} \hat{Y}_{Mean}(i,k,u)\hat{Y}_{Mean}(i,k,u') \\
\times &\left[\exp\left\{\text{Cov}[W(i,k,u) - \widehat{W}^\star(i,k,u), W(i,k,u') - \widehat{W}^\star(i,k,u'))\right\} - 1\right], \quad (18)
\end{aligned}$$

where $\text{Cov}[W(i,k,u) - \widehat{W}^\star(i,k,u), W(i,k,u') - \widehat{W}^\star(i,k,u')]$ is computed in Appendix A. (17) and (18) can be generalized in a straightforward way to expressions for the mean and variance of an arrival (or incurral) year reserve and the total reserve. The formulas in this subsection generalize the expressions from England & Verrall (2002, Section 7.7) to the framework of lognormal mixed models.

3.2 Bayesian approach: towards a full predictive distribution

Note that the formulas (16), (17) and (18) within the likelihood context require some programming using a statistical software package (or spreadsheet) and possibly are subject of discussion. For instance, they don't take the uncertainty into account that is introduced by replacing the variance components with their estimates obtained via ML or REML. Moreover, the likelihood approach only provides an estimate of the second moment of the distribution of the reserves and not its full predictive distribution. In light of these remarks, a Bayesian analysis of the proposed lognormal mixed models for claims reserving is useful. Bayesian statistics already turned out to be helpful in loss reserving with run-off triangles, as discussed in e.g. de Alba (2002), Ntzoufras & Dellaportas (2002) and England & Verrall (2002). We refer to the statistical and actuarial literature for an introduction to Bayesian statistics and their applications in actuarial statistics.

The Bayesian approach treats all unknown parameters in the lognormal mixed model as random variables. Our distributional assumptions (see Section 2) and prior specifications are summarized below:

$$\begin{aligned}
\boldsymbol{W}(i,k)|\boldsymbol{b}(i,k) &\sim N(\boldsymbol{X}(i,k)\boldsymbol{\beta}+\boldsymbol{Z}(i,k)\boldsymbol{b}(i,k),\boldsymbol{\Sigma}(i,k)),\\
\boldsymbol{b}(i,k) &\sim N(0,\boldsymbol{D}),\\
\boldsymbol{\beta} &\sim N(0,\boldsymbol{F}),\\
\boldsymbol{D} &\sim \text{Inv-Wishart}_q(\boldsymbol{B}),
\end{aligned} \tag{19}$$

where $\boldsymbol{F}$ is a diagonal matrix with large entries and $\boldsymbol{B}$ is a matrix with the same dimensions as $\boldsymbol{D}$. In the example in Section 4, $\boldsymbol{\Sigma}(i,k)=\text{diag}(\sigma_\epsilon^2)$ with prior $\sigma_\epsilon^2\sim$ Inv-Gamma(a,b).

To sample from the relevant posterior and predictive distributions, the Gibbs sampling scheme is used. The involved full conditionals are given by:

$$\begin{aligned}
[\boldsymbol{\beta}|.] &\sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})\\
\text{where } \boldsymbol{\mu}_{\boldsymbol{\beta}} &= \left(\sum_{i,k}\boldsymbol{X}(i,k)'\boldsymbol{\Sigma}(i,k)^{-1}\boldsymbol{X}(i,k)+\boldsymbol{F}^{-1}\right)^{-1}\\
&\times \sum_{i,k}\boldsymbol{X}(i,k)'\boldsymbol{\Sigma}(i,k)^{-1}(\boldsymbol{W}(i,k)-\boldsymbol{Z}(i,k)\boldsymbol{b}(i,k))\\
\text{and } \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \left(\sum_{i,k}\boldsymbol{X}(i,k)'\boldsymbol{\Sigma}(i,k)^{-1}\boldsymbol{X}(i,k)+\boldsymbol{F}^{-1}\right)^{-1},
\end{aligned} \tag{20}$$

next to,

$$\begin{aligned}
[\boldsymbol{b}(i,k)|.] &\sim N(\boldsymbol{\mu}_{\boldsymbol{b}(i,k)},\boldsymbol{\Sigma}_{\boldsymbol{b}(i,k)}) \text{ where}\\
\boldsymbol{\mu}_{\boldsymbol{b}(i,k)} &= \left(\boldsymbol{Z}(i,k)'\boldsymbol{\Sigma}(i,k)^{-1}\boldsymbol{Z}(i,k)+\boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{Z}(i,k)'\boldsymbol{\Sigma}(i,k)^{-1}(\boldsymbol{W}(i,k)-\boldsymbol{X}(i,k)\boldsymbol{\beta})\\
\text{and } \boldsymbol{\Sigma}(i,k) &= \left(\boldsymbol{Z}(i,k)'\boldsymbol{\Sigma}(i,k)^{-1}\boldsymbol{Z}(i,k)+\boldsymbol{D}^{-1}\right)^{-1},
\end{aligned} \tag{21}$$

and with N the total number of claims in our dataset,

$$[\boldsymbol{D}|.] \sim \text{Inv-Wishart}_{q+N}\left(\boldsymbol{B} + \sum_{i,k} \boldsymbol{b}(i,k)\boldsymbol{b}(i,k)'\right), \tag{22}$$

$$[\sigma_\epsilon^2|.] \sim \text{Inv-Gamma}\left(a + \frac{1}{2}\sum_{i,k} n_{ik}, b + \frac{1}{2}\sum_{i,k} \boldsymbol{W}(i,k)'\boldsymbol{W}(i,k)\right). \tag{23}$$

To perform the simulations, to check convergence of the chains and compute posterior summaries, we used the WinBUGS software. Further details are discussed in Section 4.

4 Case study

To illustrate the use of mixed models in claims reserving, a data set from a Belgian reinsurance consultant is analyzed. We first present the characteristics of the data at hand. A mixed model related to the lognormal regression model in (11) is fitted to the log-transformed data. Predictions from a likelihood analysis are obtained with Proc Mixed in SAS. A Bayesian analysis of the model leads to the full predictive distribution of the reserves and is implemented using WinBUGS. Section 4.3 concludes the case study with a discussion of the results.

4.1 Presentation of the data

The data set consists of 915 claims. The arrival (or incurral) years of the available claims vary between 1986 and 2001. Their development is followed up to 2002, unless the claim is settled earlier. For every individual payment the corresponding arrival, development and calendar year is known.

Instead of working with the complete data set, we consider a subset of 338 claims, consisting of the claims from the first 8 arrival years in the original data set. In Table 2 this subset is summarized as a classical run-off triangle with cumulative payments. In the direction of the arrival years, 1 corresponds with 1986 and 8 with 1993. Due to the choice of our subset, the lower triangle (in bold) is known and can be compared with predictions obtained via classical techniques as well as with the predictions from a lognormal mixed model for loss reserving. The latter are obtained on the scale of individual claims, but can be aggregated afterwards. For the classical technique, a model within the lognormal framework is chosen, since this enables pertinent comparisons with the fits and predictions from the lognormal mixed model. In the sequel, the results from model (9), with chain-ladder type structure for the mean, are considered as a benchmark for the results on the level of aggregate data.

Arrival	*Development Year*								
Year	1	2	3	4	5	6	7	8	9
1	19,769	1,036,536	2,926,089	3,208,614	3,710,362	3,978,786	4,429,728	4,975,137	5,348,813
2	2,531	107,813	377,475	514,688	1,106,704	1,776,792	2,201,502	2,509,058	**2,579,698**
3	23,019	88,497	432,258	716,667	1,250,396	1,623,619	2,708,759	**3,357,284**	**3,738,158**
4	495	199,119	821,879	1,275,476	1,753,482	2,156,416	**2,824,184**	**3,362,437**	**3,594,122**
5	1,116	176,504	400,207	1,037,350	2,150,087	**4,548,049**	**4,966,763**	**5,334,399**	**6,328,420**
6	3,801	134,170	929,088	1,223,499	**1,699,767**	**2,426,058**	**2,907,379**	**3,257,116**	**3,540,434**
7	22,408	331,294	960,434	**1,379,776**	**1,941,757**	**2,150,194**	**4,044,069**	**4,760,987**	**5,250,504**
8	14,246	487,661	**944,422**	**1.645,343**	**1,990,303**	**2,861,664**	**3,041,089**	**3,648,400**	**4,116,567**

Table 2: *Summary of the considered data as a classical run-off triangle with cumulative payments. Lower triangle in bold: these data have to be predicted.*

Figure 1 illustrates the extensive data set, underlying Table 2, by plotting a random selection of individual incremental (left panel) and cumulative (right panel) payment profiles until $min(\text{DY of Settlement},\ 9-\text{AY}+1)$ (where AY stands for 'arrival year' and DY for the 'development year' of the claim). To avoid problems with zero or negative payments in a lognormal model, our analysis considers cumulative data. Figure 2 shows boxplots of the log-transformed cumulative data over the available development years, for the data underlying the upper triangle in Table 2.

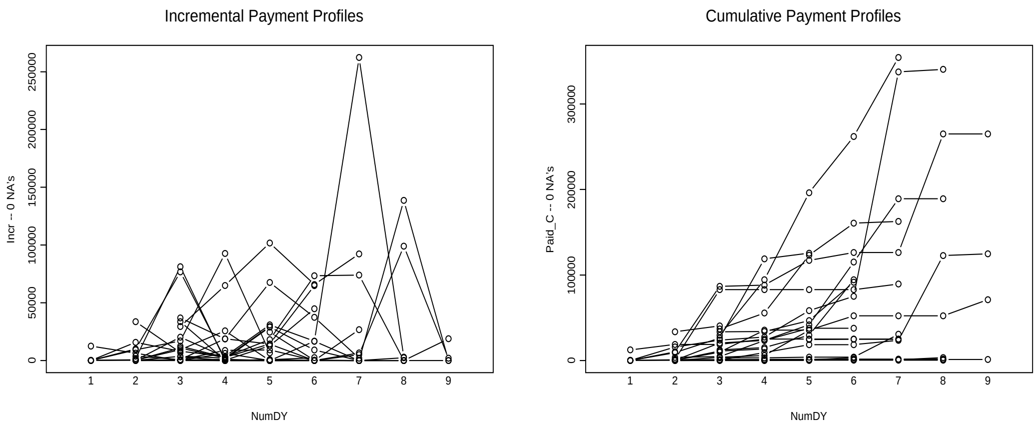

Figure 1: *Incremental (left) and cumulative (right) payment profiles over development period; randomly selected claims from the data set summarized in Table 2 (upper triangle).*

4.2 Numerical results

4.2.1 Lognormal chain-ladder model

The results of a Bayesian analysis of the lognormal regression model with chain-ladder type structure for the mean, as in (9), are given in Table 7. These are obtained with

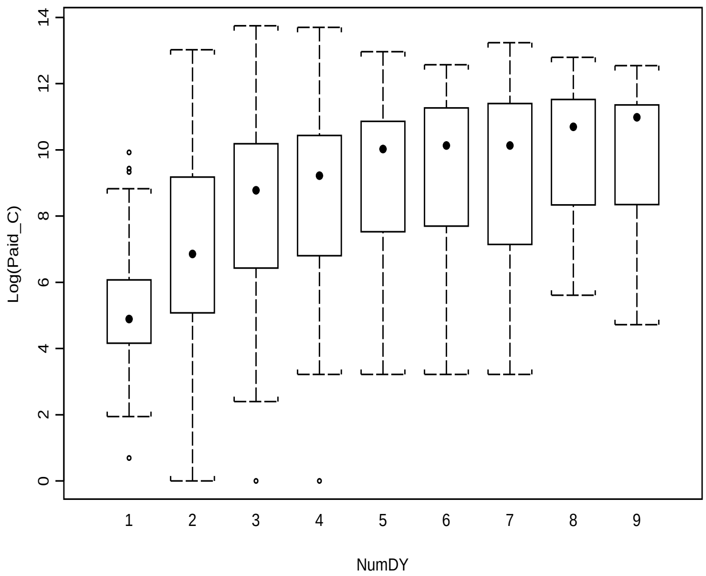

Figure 2: *Boxplots of the logarithm of cumulative payments per development year, individual data underlying the upper triangle in Table 2. • indicates median.*

WinBUGS. Within this Bayesian framework, prior specifications for the regression parameters and variance component are non-informative and similar to those discussed in Section 3.2. In this way, the values shown in the column 'Mean' in Table 7 are close to the predictions from a likelihood-based analysis (not displayed here), which can be obtained with any software package for linear regression. The results for the aggregate triangle are summarized here to enable comparisons with the predictions from mixed models on the level of arrival year or total reserves. Compared to the models presented in this paper, the classical techniques don't lead directly to predictions for an individual payment profile. However, practitioners often use the development factors from the deterministic chain-ladder to predict individual profiles. In Figure 4 the results of this ad-hoc technique are compared with the mixed model predictions for individual claims. In this figure, plots 4.3 and 4.7, respectively, compare the chain-ladder individual predictions with the real observed payments, on the logarithmic and the original scale, respectively. Plots 4.4 and 4.5 are on the original scale and compare the chain-ladder predictions with the mixed model predictions from the second (Mean) and the first (Median) formulas in (16). Further discussion of Figure 4 is postponed to Section 4.3.

4.2.2 Lognormal mixed models

Inspired by the lognormal regression models for classical run-off triangles, a whole scala of lognormal mixed models is available for claims reserving. Table 3 contains the specification of the fixed and random effects used in our analysis. Calyear is a continuous variable in the direction of calendar years and equals $\text{AY} + \text{DY} - 2$. The fixed effects' structure is inspired by (11) and the choice of the random effects by Figure 2 and a similar plot with the individual profiles of the log-transformed cumulative data. Using this specification, the design matrix $\boldsymbol{X}(i,k)$ for the fixed effects is built up as follows: one row per observation for the k^{th} claim from arrival year i, 0/1's in the columns related to the arrival year effects (where 1 indicates that the claim is from that specific arrival year) and the observed values of Calyear, DY and log (DY) in the remaining columns. The design matrix $\boldsymbol{Z}(i,k)$ is constructed in an analogous way: one row per observation for claim (i,k), a column consisting of 1's for the intercept and the observed values of DY and log (DY) in the remaining columns. In this way, claim-specific intercepts and slopes for DY and log (DY) are modelled.

The covariance matrix $\boldsymbol{D}$ of the random effects is not forced to satisfy any structural assumptions. The residual terms are modelled independently, thus $\boldsymbol{\Sigma}(i,k)$ is diagonal. Since we are dealing with cumulative data, an AR(1) structure for $\boldsymbol{\Sigma}(i,k)$ could be suggested. However, the diagonal structure came out as the preferred choice of a comparison between the empirical variance function (obtained by taking the average of the squared ordinary least squares residuals per development year) and the fitted variance function, a technique suggested by Verbeke & Molenberghs (2002). Moreover, we obtained better predictions with the diagonal residual matrix.

The suggested mixed model is implemented both in a frequentist and a Bayesian way. Recall that the data consist of individual cumulative payment profiles from the start of the development until the settlement of the claim. At first (cfr. results in Table 5 and 8) the year of settlement of the claim was taken as *a priori* information in our predictions. Of course, in practice this is less realistic and secondly the modelling of the settlement of a claim is included in the WinBUGS analysis (cfr. results in Table 9). This is done by introducing a 0/1 indicator variable for every outstanding payment

$$\begin{aligned} Z(i,k,u) &\sim Bern(1-p(u)), \\ Z(i,k,u) &= \begin{cases} 0 \text{ when last payment done in DY } u, \\ 1 \text{ otherwise.} \end{cases} \end{aligned} \tag{24}$$

Appropriate multiplication of these indicator variables with the simulations from the posterior predictive distribution of the cumulative payments allows to model the settlement of a claim. $p(j)$ $(j = 1, \ldots, 9)$ is the probability that a claim settles in DY j and is estimated by its empirical analogue, based on the complete data set of 915 claims. The estimated probabilities are given in Table 6.

Fixed Effects	Random Effects
$\alpha_1, \ldots, \alpha_8$	Intercept
Calyear	DY
DY, $\log(\text{DY})$	$\log(\text{DY})$

Table 3: *Mixed model specification.*

Table 4 contains the parameter estimates and their standard errors as obtained with Proc Mixed in SAS. The predicted values for the remaining cumulative payments of the reported claims were computed using the second formula in (16). The fitted values for the upper triangle are obtained as $\exp\left(\hat{W}^{\star}(i,k,j) + \frac{1}{2}\text{Var}(\hat{W}^{\star}(i,k,j) - W^{\star}(i,k,j))\right)$, where j refers to an observed payment. They are displayed in Table 5 to illustrate the fit of the mixed model on an aggregate basis. By adding up the fitted values for the upper triangle and the predicted values for the lower triangle appropriately, the results in Table 5 were obtained. Figure 3 illustrates the fitted profiles and predicted remaining payments for 6 randomly selected claims from the data set. The profiles are plotted on the log-scale, together with plus/minus one standard error.

Using a Bayesian analysis, simulated values are obtained for the full predictive distribution of an individual claim, a complete arrival year, or the total reserve. Hierarchical centering of the random intercepts and mean centering of the covariates is used, together with the prior specifications from Section 3.2. Table 8 (settlement *a priori*) and Table 9 (settlement modelled) display the results for the lower triangle in Table 2. The results in these tables are based on 140,000 simulations (to which a thinning factor of 5 is applied), after a burn-in of another 20,000 simulations. The predictive distribution of the cells in the lower triangle is summarized by its mean, median, standard deviation and MC error. Recall that the MC error or Monte Carlo standard error estimates the variability of the estimate of the mean reserve (for a discussion, see Scollnik, 2004). Other statistics, such as percentiles, can be obtained easily from WinBUGS.

4.3 Discussion

Plot 4.2 in Figure 4 illustrates that the lognormal mixed model fits the observed data – underlying the upper triangle in Table 2 – well. Plots 4.1 and 4.3 are on the logarithmic scale and show the individual predictions (for the lower triangle) obtained with the mixed model and the ad-hoc chain-ladder, respectively, against the real observed data. These plots illustrate that both techniques lead to comparable results on the log-scale. We want to add two remarks to the results obtained with the chain-ladder technique. It is important to note that the chain-ladder lacks statistical basis, in contrast to the mixed model approach which offers e.g. an estimate of the variability of the predictions. Furthermore,

Effect	Parameter	Estimate (s.e.)
Arrival Year		
Effects:	α_2	5.1335 (0.3727)
	α_3	10.3467 (0.5585)
	α_4	16.0661 (0.8410)
	α_5	21.3755 (1.1523)
	α_6	26.9522 (1.4714)
	α_7	32.9084 (1.7774)
	α_8	38.5002 (2.1154)
Calyear	γ	-5.5011 (0.3362)
Time	β_1	5.1159 (0.3609)
Log(Time)	β_2	3.7983 (0.3396)
Covariance of random effects:		
Var(b_1)	d_1	6.8411 (1.0015)
Var(b_2)	d_2	0.3695 (0.1422)
Var(b_3)	d_3	10.8311 (2.6733)
Cov(b_1,b_2)	$d_{12} = d_{21}$	0.2317 (0.2947)
Cov(b_1,b_3)	$d_{13} = d_{31}$	-3.5705 (1.3812)
Cov(b_2,b_3)	$d_{23} = d_{32}$	-1.8514 (0.5902)
Residual variance:		
Var(ϵ)	σ^2_ϵ	0.7315 (0.0492)
-2 REML log-likelihood		4260.7
AIC		4274.7

Table 4: *Parameter estimates as obtained with* Proc Mixed *analysis of the extensive data set underlying the upper triangle in Table 2.*

the success of the deterministic chain-ladder can in part be explained by the fact that it directly uses the last observed cumulative payment in every profile.

Next, the question can be asked what the predictive power of the mixed model is on the level of aggregate reserves. Tables 5, 8 and 9 illustrate that the use of individual data leads to reasonable predictions for the different cells in the lower triangle. Moreover, when the columns '% Bayes. St. Err.' and '% RMSEP' in Table 7 are compared with the corresponding columns in Tables 8 and 9, one can conclude that the mixed model performs comparably with and often even better than the stochastic chain-ladder. However, for cells in the final development years from recent arrival years (e.g. (7,8), (7,9) and (8,9)), a very large standard deviation of the predictive distribution is observed.

Note again that the predictions based on the mean of the lognormal distribution often severly overestimate the real observed payments, whereas use of the median leads to more

Arrival Year	*Development Year* 1	2	3	4	5	6	7	8	9
1	22,036	708,112	2,000,287	3,134,621	3,696,074	4,107,102	4,498,656	4,696,801	4,808,596
2	2,646	70,176	277,864	598,698	1,162,101	1,627,626	1,919,125	2,138,102	**3,084,274**
3	23,583	87,415	307,259	697,528	1,283,885	1,773,943	2,082,374	**3,034,468**	**3,116,201**
4	789	156,440	572,914	1,214,474	2,040,289	2,522,600	**3,557,169**	**4,091,425**	**4,309,365**
5	1,233	125,352	416,852	1,026,010	1,773,031	**3,346,765**	**3,886,908**	**4,207,342**	**4,474,187**
6	4,179	117,858	757,523	1,408,195	**2,817,320**	**3,536,078**	**4,075,013**	**4,317,244**	**4,500,767**
7	27,618	240,841	841,118	**2,725,891**	**4,962,650**	**6,849,153**	**8,241,505**	**9,007,422**	**9,331,285**
8	14,873	396,933	**1,927,582**	**3,091,773**	**4,087,168**	**4,815,477**	**7,677,933**	**7,656,123**	**7,705,730**

Table 5: *Fitted values for upper and predictions for lower triangle as obtained with a lognormal mixed model analysis of the extensive data set. Year of settlement taken as a priori information.*

DY	$p(j)$	DY	$p(j)$	DY	$p(j)$
1	0	4	0.2496	7	0.4637
2	0.1068	5	0.3227	8	0.5521
3	0.1809	6	0.3940	9	0.5613

Table 6: *Estimated probabilities of settlement in DY j.*

reasonable predictions. This is clearly illustrated by plots 4.6 and 4.8 of Figure 4. The predictions obtained with the first formula in (16) are also much closer to the chain-ladder predictions, than those obtained with the second formula in (16), see plots 4.4 and 4.5. The back-transformation of the predictions on the log-scale to the original scale – where the difference between the mean and the median lies in the inclusion of a variance term, see (16) – is a general problem in reserving models within the lognormal framework, see e.g. England & Verrall (2002).

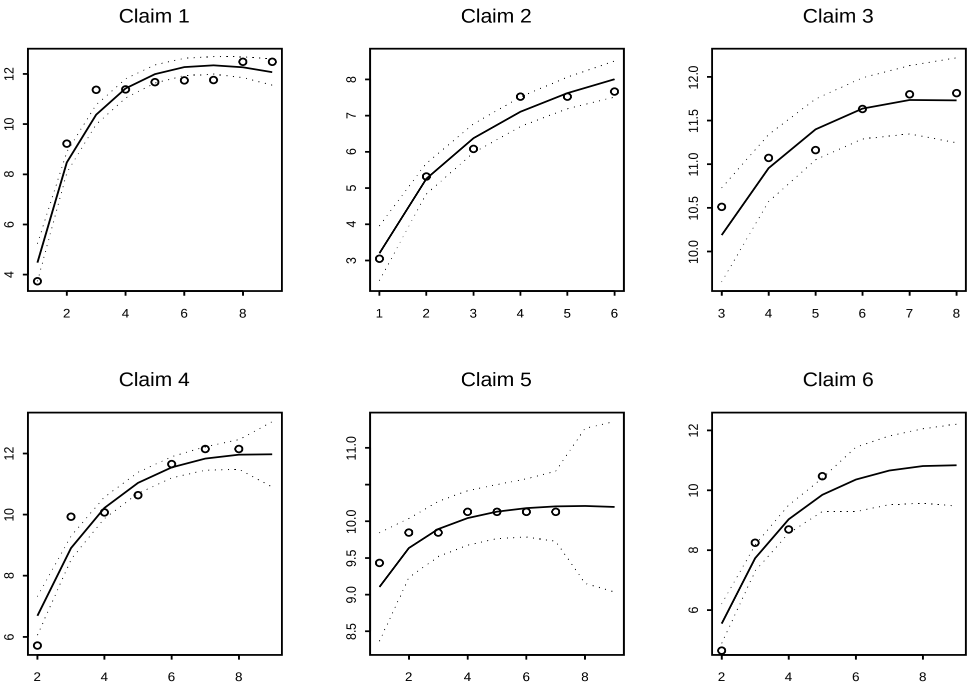

Figure 3: *Fitted and predicted payment profiles on log-scale, together with plus/minus ONE standard error. Circles are used for observed payments. Full lines give fitted and predicted profiles. Dotted lines give fits/predictions plus/minus one standard error as computed with* Proc MIXED.

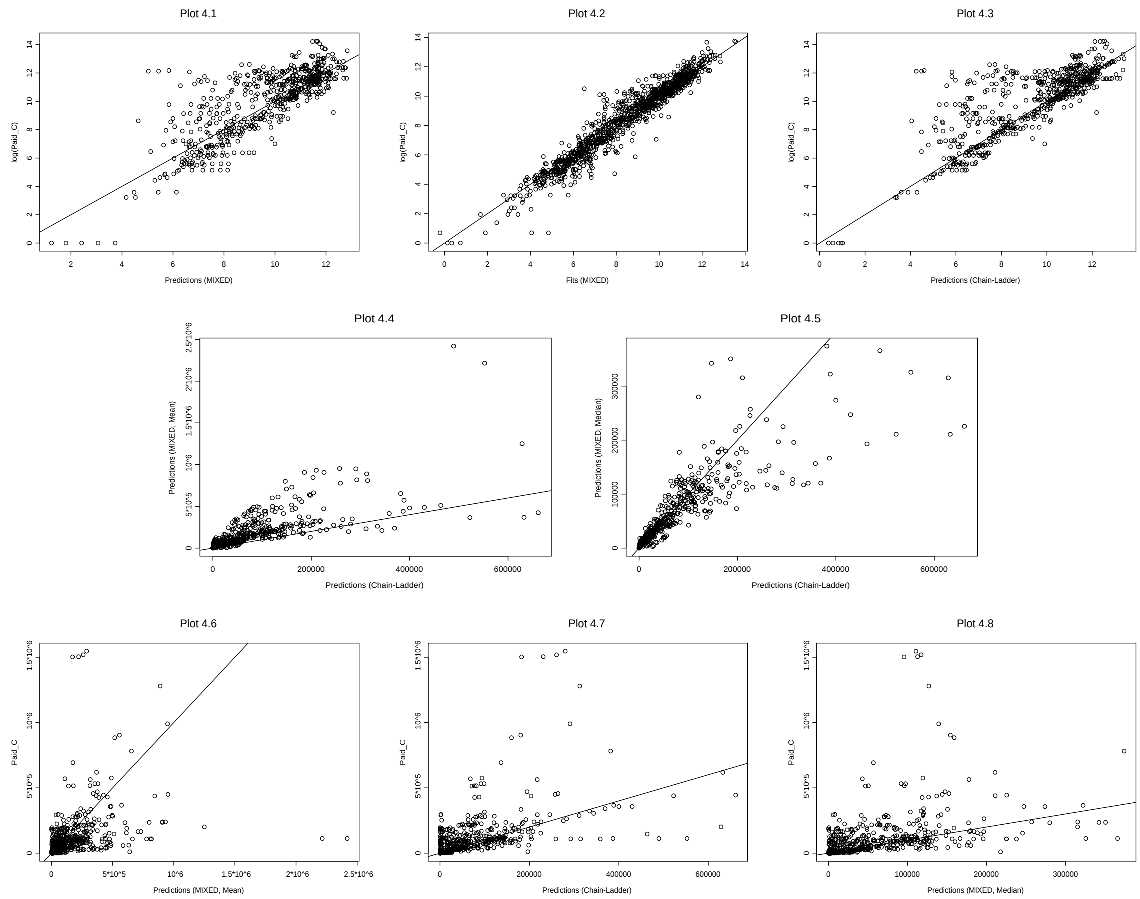

Figure 4: *Comparison between the fitted and predicted payments and the real observations. Plots in the first row are on the log-scale. Those in the second and third row are on the original scale of the payments. A line with zero intercept and slope one is added in each plot.*

Cell	Mean	St. Dev.	MC Error	Median	% Bayes. St. Err.	% RMSEP	Cell	Mean	St. Dev.	MC Error	Median	% Bayes. St. Err.	% RMSEP
(2,9)	2,890,570	1,097,152	2,998	2,673,915	37.95%	44.20%	(6,9)	5,281,017	4,062,074	13,267	4,252,943	76.92%	124.82%
(3,8)	3,476,132	1,487,329	4,077	3,115,586	42.79%	44.44%	(7,4)	2,251,030	2,141,197	5,994	1,693,554	95.12%	167.54%
(3,9)	4,054,406	2,168,533	6,099	3,534,674	53.49%	58.62%	(7,5)	4,825,450	5,253,024	15,220	3,486,996	108.86%	308.61%
(4,7)	2,871,268	1,530,306	4,203	2,557,772	53.30%	54.21%	(7,6)	6,508,785	6,449,547	20,419	4,856,408	99.09%	362.02%
(4,8)	3,401,533	1,871,671	5,264	2,982,743	55.02%	55.68%	(7,7)	8,833,172	8,429,933	28,333	6,661,597	95.43%	239.74%
(4,9)	3,807,211	3,888,859	10,545	3,295,992	102.14%	108.36%	(7,8)	10,561,739	9,929,984	34,105	8,033,604	94.02%	241.55%
(5,6)	2,741,372	959,687	2,711	2,491,472	35.01%	44.98%	(7,9)	11,857,801	11,212,943	39,637	9,033,520	94.56%	247.88%
(5,7)	3,555,032	2,008,559	5,827	3,096,994	56.50%	49.43%	(8,3)	3,237,745	5,430,145	14,549	1,938,001	167.71%	624.14%
(5,8)	4,156,925	2,455,416	7,043	3,583,686	59.07%	51.05%	(8,4)	5,066,480	6,893,491	19,383	3,322,473	136.07%	467.73%
(5,9)	4,605,369	2,832,752	8,345	3,940,136	61.51%	52.39%	(8,5)	8,696,819	10,654,973	30,010	5,885,156	122.52%	632.56%
(6,5)	2,312,128	1,721,736	4,832	1,847,352	74.47%	107.51%	(8,6)	11,104,784	12,903,307	38,991	7,704,502	116.20%	535.06%
(6,6)	3,027,390	2,272,701	6,728	2,439,669	75.07%	96.90%	(8,7)	14,397,201	16,354,112	51,467	10,121,738	113.59%	654.71%
(6,7)	4,011,684	3,120,807	9,465	3,225,342	77.79%	113.86%	(8,8)	16,840,347	18,646,491	61,021	11,958,527	110.73%	626.06%
(6,8)	4,737,237	3,591,395	11,381	3,823,356	75.81%	119.26%	(8,9)	18,675,144	20,873,632	68,390	13,267,934	111.77%	618.21%

Table 7: *Bayesian predictions based on a lognormal model with chain-ladder structure for the mean. Cell (i, j) refers to AY i and DY j in the triangle. 140,000 simulations are used, after a burn-in of 20,000 simulations. '% Bays. St. Err.' is the ratio of 'St. Dev.' and 'Mean'. '% RMSEP' is the ratio of 'RMSEP' and the real observed value for the cell (as given in Table 2). An estimate of the RMSEP is obtained from* WinBUGS *by taking the root of the mean of the distribution of the squared difference between the simulated predictive value for the cell and its real observed value, as given in Table 2.*

Cell	Mean	St. Dev.	MC Error	Median	% Bayes. St. Err.	% RMSEP	Cell	Mean	St. Dev.	MC Error	Median	% Bayes. St. Err.	% RMSEP
(2,9)	3,721,059	1,877,683	12,235	3,276,821	50.46%	85.18%	(6,9)	5,526,360	5,814,111	40,270	4,326,720	105.21%	174%
(3,8)	3,435,006	1,881,434	12,460	2,991,233	54.77%	56.09%	(7,4)	2,724,937	1,287,821	6,897	2,442,725	47.26%	134.97%
(3,9)	3,671,434	2,181,711	14,191	3,140,950	59.42%	58.39%	(7,5)	4,930,384	2,923,266	19,540	4,234,285	59.29%	215.30%
(4,7)	3,645,540	1,251,633	7,999	3,383,585	34.33%	53.01%	(7,6)	7,095,554	5,011,952	32,684	5,826,777	70.64%	327.46%
(4,8)	4,174,125	1,905,590	12,654	3,772,180	45.65%	61.60%	(7,7)	9,415,253	18,146,731	106,873	7,247,292	192.74%	467.97%
(4,9)	4,803,203	3,236,648	19,521	4,110,119	67.39 %	96.13 %	(7,8)	11,870,047	15,455,097	94,618	8,416,225	130.20%	357.32%
(5,6)	3,456,941	1,403,932	8,131	3,188,319	40.61%	39.10%	(7,9)	14,575,134	29,918,212	188,998	9,276,735	205.27%	596.85%
(5,7)	4,243,928	1,943,948	12,710	3,837,565	45.81%	41.76%	(8,3)	1,999,595	2,047,826	11,265	1,447,272	102.41%	243.93%
(5,8)	4,989,887	2,821,196	17,317	4,329,389	56.54%	53.28%	(8,4)	3,563,286	4,661,129	27,248	2,424,848	130.81%	306.34%
(5,9)	5,888,674	5,528,611	36,525	4,652,212	93.89%	87.64%	(8,5)	4,620,942	6,662,217	41,563	3,128,376	144.17%	359.88%
(6,5)	2,923,249	1,276,196	7,897	2,641,108	43.66%	104.01%	(8,6)	5,404,124	8,404,535	49,998	3,556,251	155.52%	306.84%
(6,6)	3,665,371	1,741,554	11,340	3,275,136	47.51%	88.11%	(8,7)	8,008,317	18,259,767	114,832	4,484,344	228.01%	622.26%
(6,7)	4,254,686	2,349,236	14,812	3,714,463	55.22%	93.15%	(8,8)	8,546,360	24,487,918	166,022	4,630,656	286.53%	684.49%
(6,8)	4,798,417	3,390,536	23,512	4,027,421	70.66%	114.35%	(8,9)	9,349,987	28,476,659	179,173	4,936,828	304.56%	703.34%

Table 8: *Bayesian predictions obtained from a lognormal mixed model for the extensive data set underlying Table 2. Cell (i, j) refers to AY i and DY j in the triangle. 140,000 simulations are used with a thinning factor of 5, after a burn-in of 20,000 simulations. Year of settlement taken as a priori information. '% Bays. St. Err.' is the ratio of 'St. Dev.' and 'Mean'. '% RMSEP' is the ratio of 'RMSEP' and the real observed value for the cell (as given in Table 2). An estimate of the RMSEP is obtained from* WinBUGS *by taking the root of the mean of the distribution of the squared difference between the simulated predictive value for the cell and its real observed value, as given in Table 2.*

Cell	Mean	St. Dev.	MC Error	Median	% Bayes. St. Err.	% RMSEP	Cell	Mean	St. Dev.	MC Error	Median	% Bayes. St. Err.	% RMSEP
(2,9)	3,708,159	1,865,442	12,011	3,270,076	50.31%	84.51%	(6,9)	5,602,181	7,830,627	55,223	4,032,780	139.78%	228.71%
(3,8)	3,424,877	1,818,933	11,895	2,984,185	53.11%	54.22%	(7,4)	2,736,579	1,350,757	8,145	2,442,515	49.36%	138.76%
(3,9)	3,580,979	2,178,356	15,949	3,040,418	60.83%	58.43%	(7,5)	4,380,823	2,630,209	16,551	3,771,714	60.04%	184.73%
(4,7)	3,646,895	1,232,737	8,299	3,402,781	33.80%	52.48%	(7,6)	6,283,996	4,929,445	30,708	5,117,633	78.44%	299.20%
(4,8)	4,004,876	1,659,659	10,981	3,643,519	41.44%	52.93%	(7,7)	8,216,689	7,728,357	52,157	6,294,188	94.06%	217.18%
(4,9)	4,538,662	3,004,232	19,573	3,902,436	66.19%	87.62%	(7,8)	10,482,090	14,717,253	101,995	7,337,856	140.40%	331.66%
(5,6)	3,447,314	1,278,046	8,146	3,186,867	37.07%	37.09%	(7,9)	13,409,490	31,011,116	196,189	8,021,598	231.26%	610.73%
(5,7)	4,023,425	1,716,983	11,518	3,647,260	42.67%	39.44%	(8,3)	2,008,680	2,205,517	12,921	1,463,852	109.80%	259.30%
(5,8)	4,784,864	2,804,755	17,349	4,132,994	58.62%	53.58%	(8,4)	4,394,146	9,236,956	55,286	2,680,444	210.21%	585.73%
(5,9)	5,738,788	5,181,029	35,386	4,505,410	90.28%	82.40%	(8,5)	7,178,209	16,763,572	113,483	3,934,932	233.53%	881.67%
(6,5)	2,848,404	1,161,298	6,991	2,598,159	40.77%	96.10%	(8,6)	9,216,117	22,449,661	135,745	4,804,502	243.59%	815.32%
(6,6)	3,466,732	1,759,291	10,780	3,074,718	50.75%	84.25%	(8,7)	10,355,836	26,556,155	181,599	5,219,159	256.44%	905.77%
(6,7)	4,018,304	2,485,042	15,746	3,433,040	61.84%	93.63%	(8,8)	11,277,208	32,632,283	197,276	5,479,716	289.36%	918.54%
(6,8)	4,680,612	4,927,990	31,014	3,730,880	105.29%	157.48%	(8,9)	12,781,919	133,210,650	804,203	5,700,636	1042.18%	3242.80%

Table 9: *Bayesian predictions as obtained with a lognormal mixed model for extensive data set underlying Table 2. Cell (i,j) refers to AY i and DY j in the triangle. 140,000 simulations with a thinning factor of 5, after a burn-in of 20,000 simulations. Settlement modelled explicitly. '% Bays. St. Err.' is the ratio of 'St. Dev.' and 'Mean'. '% RMSEP' is the ratio of 'RMSEP' and the real observed value for the cell (as given in Table 2). An estimate of the RMSEP is obtained from* WinBUGS *by taking the root of the mean of the distribution of the squared difference between the simulated predictive value for the cell and its real observed value, as given in Table 2.*

5 Conclusions

This paper introduces the use of mixed models in claims reserving. Both a likelihood-based as well as a Bayesian implementation of the lognormal mixed models are discussed. Within the likelihood approach, expressions for the mean and variance of an individual, arrival year and the total reserve are explained. In this way the expressions in England & Verrall (2002) are generalized to the framework of lognormal mixed models. The Gibbs sampling scheme for the Bayesian analysis is set up. This new approach to claims reserving on an individual data basis is illustrated with a case study from practice. Further work in this direction should focus on the use of generalized linear and generalized additive mixed models for loss reserving and the appropriate modelling of zeros. A stochastic discounting process can also be included.

Acknowledgements

The authors would like to thank an Associate Editor and two anonymous referees for their comments on an earlier version of this paper, which lead to a considerable improvement of the presentation of our work. Illustrative SAS or WinBUGS code can be obtained on request from the first author.

A Covariance Expression in (18)

In this section we describe how an expression for the covariance in (18) can be derived. In the sequel it is assumed that the residual terms are modelled independently (with $\Sigma(i,k) = \text{diag}(\sigma_\epsilon^2)$), but the formulas can be generalized in a straightforward way.

$$\begin{aligned} & \text{Cov}[W(i,k,u) - \widehat{W}^\star(i,k,u), W(i,k,u') - \widehat{W}^\star(i,k,u')] \\ = \; & \text{Cov}[W(i,k,u), W(i,k,u')] - \text{Cov}[W(i,k,u), \widehat{W}^\star(i,k,u')] \\ - \; & \text{Cov}[\widehat{W}^\star(i,k,u), W(i,k,u')] + \text{Cov}[\widehat{W}^\star(i,k,u), \widehat{W}^\star(i,k,u')]. \end{aligned} \tag{25}$$

The first term in this expression is given by, with $\delta_{u,u'} = 0$ if $u \neq u'$ and 1 if $u = u'$:

$$\text{Cov}[W(i,k,u), W(i,k,u')] = \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{z}(i,k,u') + \delta_{u,u'}\sigma_\epsilon^2. \tag{26}$$

Some matrix calculations lead to

$$\begin{aligned} \text{Cov}[W(i,k,u), \widehat{W}^\star(i,k,u')] \; = \; & \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{X}(i,k)(\sum_h \boldsymbol{X}_h'\boldsymbol{V}_h^{-1}\boldsymbol{X}_h)^{-1} \\ \times \; & \{\boldsymbol{x}(i,k,u')' - \boldsymbol{z}(i,k,u')'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{X}(i,k)\}' \\ + \; & \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{Z}(i,k)\boldsymbol{D}\boldsymbol{z}(i,k,u'), \end{aligned} \tag{27}$$

and

$$\begin{aligned}
& \mathrm{Cov}[\hat{W}^\star(i,k,u), W(i,k,u')] \\
= \; & \{\boldsymbol{x}(i,k,u)' - \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{X}(i,k)\}(\sum_h \boldsymbol{X}_h'\boldsymbol{V}_h^{-1}\boldsymbol{X}_h)^{-1} \\
\times \; & \boldsymbol{X}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{Z}(i,k)\boldsymbol{D}\boldsymbol{z}(i,k,u') \\
+ \; & \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\boldsymbol{Z}(i,k)\boldsymbol{D}\boldsymbol{z}(i,k,u'). \qquad (28)
\end{aligned}$$

Furthermore

$$\begin{aligned}
& \mathrm{Cov}[\widehat{W}^\star(i,k,u), \widehat{W}^\star(i,k,u')] \\
= \; & \mathrm{Cov}[\boldsymbol{x}(i,k,u)'\widehat{\boldsymbol{\beta}}, \boldsymbol{x}(i,k,u')'\widehat{\boldsymbol{\beta}}] + \mathrm{Cov}[\boldsymbol{z}(i,k,u)'\widehat{\boldsymbol{b}}(i,k), \boldsymbol{z}(i,k,u')'\widehat{\boldsymbol{b}}(i,k)],
\end{aligned}$$

because $\mathrm{Cov}[\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}(i,k)] = 0$. Now use

$$\begin{aligned}
& \mathrm{Var}[\widehat{\boldsymbol{b}}(i,k)] \\
= \; & \boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1}\{\boldsymbol{V}(i,k) - \boldsymbol{X}(i,k)(\sum_h \boldsymbol{X}_h'\boldsymbol{V}_h^{-1}\boldsymbol{X}_h)^{-1}\boldsymbol{X}(i,k)'\}\boldsymbol{V}(i,k)^{-1}\boldsymbol{Z}(i,k)\boldsymbol{D},
\end{aligned}$$

and

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\sum_h \boldsymbol{X}_h'\boldsymbol{V}_h^{-1}\boldsymbol{X}_h)^{-1}.$$

From which we conclude

$$\begin{aligned}
& \mathrm{Cov}[\widehat{W}^\star(i,k,u), \widehat{W}^\star(i,k,u')] \\
= \; & \boldsymbol{x}(i,k,u)'(\sum_h \boldsymbol{X}_h'\boldsymbol{V}_h^{-1}\boldsymbol{X}_h)^{-1}\boldsymbol{x}(i,k,u') + \boldsymbol{z}(i,k,u)'\boldsymbol{D}\boldsymbol{Z}(i,k)'\boldsymbol{V}(i,k)^{-1} \\
\times \; & \{\boldsymbol{V}(i,k) - \boldsymbol{X}(i,k)(\sum_h \boldsymbol{X}_h\boldsymbol{V}_h^{-1}\boldsymbol{X}_h)^{-1}\boldsymbol{X}(i,k)'\}\boldsymbol{V}(i,k)^{-1}\boldsymbol{Z}(i,k)\boldsymbol{D}\boldsymbol{z}(i,k,u')'. (29)
\end{aligned}$$

Combining (26), (27), (28) and (29) into (25) then leads to an expression for the covariance term in (18).

References

[1] Barnett, G. & Zehnwirth, B. (1998). *Best estimates for reserves.* Casualty Actuarial Science Forum, 1-54.

[2] de Alba, E. (2002). *Bayesian estimation of outstanding claims reserves.* North American Actuarial Journal, 6(4), 1-20.

[3] Demidenko, E. (2004). *Mixed Models: Theory and Applications.* Wiley Series in Probability and Statistics, Hoboken, New Jersey.

[4] De Vylder, F. & Goovaerts, M.J. (1979). *Proceedings of the first meeting of the contact group "Actuarial Science".* KU Leuven, nr. 7904B, wettelijk depot: D/1979/23761/5.

[5] Diggle, P.J., Heagerty, P., Liang, K.Y. & Zeger, S.L. (2002). *Analysis of longitudinal data.* Oxford University Press, Oxford.

[6] Doray, L.G. (1996). *UMVUE of the IBNR reserve in a lognormal linear regression model.* Insurance: Mathematics and Economics, 18(1), 43-57.

[7] England, P.D. & Verrall, R.J. (2002). *Stochastic claims reserving in general insurance.* British Actuarial Journal, 8, 443-544.

[8] Frees, E.W., Young, V.R. and Luo, Y. (1999). *A longitudinal data analysis interpretation of credibility models.* Insurance: Mathematics and Economics, 24(3), 229-247.

[9] Frees, E.W., Young, V.R. and Luo Y. (2001). *Case studies using panel data models.* North American Actuarial Journal, 5(4), 24-42.

[10] Haastrup, S. & Arjas, E. (1996). *Claims reserving in continuous time: a nonparametric Bayesian approach.* ASTIN Bulletin, 26(2), 139-164.

[11] Kremer, E. (1982). *IBNR claims and the two way model of ANOVA.* Scandinavian Actuarial Journal, 47-55.

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

[13] Laird, N.M. & Ware, J.H. (1982). *Random-effects models for longitudinal data.* Biometrics, 38, 963-974.

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

[15] Ntzoufras, I. & Dellaportas, P. (2002). *Bayesian modelling of outstanding lialibilities incorporating claim count uncertainty.* North American Actuarial Journal, 6, 113-128.

[16] Scollnik, D.P.M. (2004). *Bayesian reserving models inspired by chain-ladder methods and implemented using WinBUGS.* ARCH, 2004(2).

[16] Taylor, G. & Campbell, M. (2002). *Statistical case estimation.* Research paper number 104, The University of Melbourne, Australia.
available online at http://www.economics.unimelb.edu.au/actwww/html/no104.pdf

[17] Taylor, G., McGuire, G. & Greenfield, A. (2003). *Loss reserving: past, present and future.* Research paper number 109, The University of Melbourne, Australia.
available online at http://www.economics.unimelb.edu.au/actwww/html/no109.pdf

[18] Verbeke, G. & Molenberghs, G. (1997) *Linear mixed models in practice: a SAS oriented approach.* Lecture notes in statistics, Springer New York.

[19] Verbeke, G. & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* Springer Series in Statistics, Springer New York.

[20] Zehnwirth, B. (1985). *ICRFS version 4 manual and users guide.* Benhar Nominees PTy Ltd, Turramurra, NSW, Australia.