

LOGNORMAL MIXED MODELS FOR REPORTED CLAIMS RESERVES

Katrien Antonio,^{*} Jan Beirlant,[†] Tom Hoedemakers,[‡] and Robert Verlaak[§]

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. This contribution explores 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 with, for example, confidence regions. 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 simulation from the complete predictive distribution of the reserves and the calculation of various risk measures. The paper ends with an illustration of the suggested techniques on a data set from practice, consisting of Belgian automotive third-party liability claims. The results for the mixed-model analysis are compared with those obtained from traditional claims-reserving techniques for run-off triangles. For the data under consideration, the lognormal mixed model fits the observed individual data well. It leads to individual predictions comparable to those obtained by applying chain-ladder development factors to individual data. Concerning the predictive power on the aggregate level, the mixed model leads to reasonable predictions and performs comparable to and often better than the stochastic chain ladder for aggregate data.

1. INTRODUCTION

Claims originating in a particular year often cannot be finalized in the same year. Many causes for delay in 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 are already incurred 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 toward 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 and 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 $i, j = 1, \dots, t$) 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

^{*} Katrien Antonio is a PhD student in the University Center for Statistics, KU Leuven, W. de Croijlaan 54, 3001 Heverlee, Belgium, katrien.antonio@econ.kuleuven.be.

[†] Jan Beirlant is a full professor in the University Center for Statistics, KU Leuven, W. de Croijlaan 54, 3001 Heverlee, Belgium.

[‡] Tom Hoedemakers is a postdoctoral researcher in the Department of Applied Economics, Naamsestraat 69, 3000 Leuven, Belgium.

[§] Robert Verlaak is an actuary at AON Re Belgium, Van Nieuwenhuyselaan 2, 1160 Brussels, Belgium.

Table 1
Random Variables in a Run-off Triangle

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

(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 for aggregate data is to complete this run-off triangle to a square by predicting future payments.

The present literature on loss reserving mainly describes techniques based on summary triangles like Table 1. However, some authors recently suggested leaving the track of aggregate claim figures. To illustrate this statement we quote England and 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 and Campbell (2002) and Taylor, McGuire, and Greenfield (2003, pp. 21–22) 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.” 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 and Arjas (1996) suggested modeling the data from individual claims in a nonparametric Bayesian way with the occurrence and development of claims modeled 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 in 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, McGuire, and Greenfield 2003):

1. A unique reference to denote each claim in the data set
2. A record for each payment made for a particular claim and (if available) each change in the company’s estimate of the incurred loss
3. The arrival and reporting year for each claim, the development, and the calendar year to which a payment belongs
4. (If available) information concerning specific features of the policyholder (e.g., age and gender).

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, which is 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 and Ware (1982). In this paper we explain how individual record data sets can be modeled in the context of mixed models and how these models lead to forecasts for future payments for the reported but not completely settled claims. It is important to point out that the logarithm of the individual data is modeled 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, Young, and Luo (1999, 2001). This text presents another actuarial domain, namely, loss reserving, where the models can be useful.

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

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. Then we describe how loss reserving based on individual record data sets can be performed within the framework of linear mixed models, in both a likelihood and a Bayesian way. The paper ends with the illustration of the presented techniques on the case study data set.

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 and 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. First, 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 that allow the number of measurements (here: payments or loss ratios) and times of observations to vary among subjects (here: claims). Second, methods are needed that enable modeling the dependencies among the observations on a certain subject/claim. Imagine, for instance, that individual cumulative payments are modeled; then observations on the same claim cannot 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 modeled 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. 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 specified as

$$Y_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, N. \quad (2.1)$$

$\boldsymbol{\beta}$ ($p \times 1$) gives the p fixed-effects parameters. These are fixed, but unknown, regression parameters, common to all subjects. \mathbf{b}_i ($q \times 1$) is the vector with the random effects 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 $\boldsymbol{\beta}$. X_i ($n_i \times p$) and Z_i ($n_i \times q$) are the design matrices for the p fixed and q random effects, and $\boldsymbol{\varepsilon}_i$ ($n_i \times 1$) contains the residual components for subject i . Independence between subjects is assumed. Here \mathbf{b}_i and $\boldsymbol{\varepsilon}_i$ also are assumed to be independent, and we follow the traditional assumption that they are normally distributed with mean vector $\mathbf{0}$ and covariance matrices, say, \mathbf{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 Verbeke and Molenberghs (1997, 2000). It is easy to see that Y_i then has a marginal normal distribution with mean $X_i\boldsymbol{\beta}$ and covariance matrix $V_i = \text{Var}(Y_i)$, given by

$$V_i = Z_i\mathbf{D}Z_i' + \boldsymbol{\Sigma}_i. \quad (2.2)$$

In this interpretation it becomes clear that the fixed effects enter only the mean $E[Y_i]$, 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,

$$\begin{aligned} Y_i|\mathbf{b}_i &\sim N(X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i, \boldsymbol{\Sigma}_i), \\ \mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}), \end{aligned} \quad (2.3)$$

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

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

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^N X_i'V_i^{-1}X_i \right)^{-1} \sum_{i=1}^N X_i'V_i^{-1}Y_i. \quad (2.4)$$

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, $\mathbf{b}_i|Y_i$, is used. Conditional on $\boldsymbol{\alpha}$, we have

$$\hat{\mathbf{b}}_i = \mathbf{D}Z_i'V_i^{-1}(Y_i - X_i\hat{\boldsymbol{\beta}}), \quad (2.5)$$

which can be proven to be the Best Linear Unbiased Predictor (BLUP) of \mathbf{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

$$L_1(\boldsymbol{\alpha}; \mathbf{y}_1, \dots, \mathbf{y}_N) = c_1 - \frac{1}{2} \sum_{i=1}^N \log|V_i| - \frac{1}{2} \sum_{i=1}^N \mathbf{r}_i'V_i^{-1}\mathbf{r}_i, \quad (2.6)$$

$$L_2(\boldsymbol{\alpha}; \mathbf{y}_1, \dots, \mathbf{y}_N) = c_2 - \frac{1}{2} \sum_{i=1}^N \log|V_i| - \frac{1}{2} \sum_{i=1}^N \mathbf{r}_i'V_i^{-1}\mathbf{r}_i - \frac{1}{2} \sum_{i=1}^N \log|X_i'V_i^{-1}X_i|, \quad (2.7)$$

where $\mathbf{r}_i = \mathbf{y}_i - X_i(\sum_{i=1}^N X_i'V_i^{-1}X_i)^{-1}(\sum_{i=1}^N X_i'V_i^{-1}\mathbf{y}_i)$ and c_1, c_2 are appropriate constants. Equations (2.6) and (2.7) are maximized using iterative numerical techniques such as Fisher scoring or Newton-Raphson (for full details, see Demidenko 2004). In equations (2.4) and (2.5) 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 \mathbf{b}_i . For inference regarding the fixed and random effects and the variance components, appropriate likelihood ratio and Wald tests are explained in Verbeke and Molenberghs (2000).

The predictor for the conditional expectation $Y_i^* := E[Y_i|\mathbf{b}_i] = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i$ is obtained from equations (2.4) and (2.5), namely,

$$\begin{aligned}
\hat{Y}_i^* &= X_i \hat{\beta} + Z_i \hat{b}_i \\
&= X_i \hat{\beta} + Z_i D Z_i' V_i^{-1} (Y_i - X_i \hat{\beta}) \\
&= (I_{n_i} - Z_i D Z_i' V_i^{-1}) X_i \hat{\beta} + Z_i D Z_i' V_i^{-1} Y_i \\
&= \Sigma_i V_i^{-1} X_i \hat{\beta} + (I_{n_i} - \Sigma_i V_i^{-1}) Y_i.
\end{aligned}$$

Note that this expression can be interpreted as a *credibility* predictor, because it is a weighted average of $X_i \hat{\beta}$ (related to the whole database) and Y_i (related to subject i). The credibility weights are $\Sigma_i V_i^{-1}$ and $I_{n_i} - \Sigma_i V_i^{-1}$, which implies that $X_i \hat{\beta}$ gets much weight if the residual variability is “large” in comparison with the total variability. \hat{Y}_i^* is the BLUP of Y_i^* . If the residual terms are modeled independently (thus Σ_i diagonal for every i), $\hat{Y}_{i,n_i+1}^* = x'_{i,n_i+1} \hat{\beta} + z'_{i,n_i+1} \hat{b}_i$ is also the BLUP for $Y_{i,n_i+1} = x'_{i,n_i+1} \beta + z'_{i,n_i+1} b_i + \varepsilon_{i,n_i+1}$.

Verbeke and Molenberghs (1997) describe in a detailed way how different types of mixed models can be fitted 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 then is set up to sample from the relevant posterior and predictive distributions. More details are given in Section 3.2. WINBUGS² 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 favor the use of mixed models in a practical loss-reserving context.

3. REPORTED CLAIMS 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 us 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) modeling 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 adapted to model zero or negative incremental payments or censored data by using two-part models based on generalized linear mixed models (GLMM).

Applied to a run-off triangle like Table 1, the general lognormal regression model is given by

$$\log(Y) = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \quad (3.1)$$

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

$$\log(Y_{ij}) = \alpha_i + \beta_j + \varepsilon_{ij}, \quad (3.2)$$

¹ SAS is a commercial software package (for details see 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 www.mrc-bsu.cam.ac.uk/bugs).

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

$$\log(Y_{ij}) = \alpha_i + \sum_{l=1}^{j-1} \beta_l + \sum_{t=1}^{i+j-2} \gamma_t + \epsilon_{ij}, \quad (3.3)$$

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

$$\log(Y_{ij}) = \alpha_i + \beta \log(j) + \gamma_j + \epsilon_{ij}. \quad (3.4)$$

To set up the reserves in a run-off triangle, one has to forecast the lower triangle in Table 1, namely, Y_{ij} values with $i + j > t + 1$.

Recall the notation introduced in Section 1 to describe an individual claims 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 . $Y(i, k) = \{Y(i, k, 1), \dots, Y(i, k, n_{ik})\}'$ 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

$$\begin{aligned} W(i, k) &:= \log(Y(i, k)) = X(i, k)\beta + Z(i, k)b(i, k) + \epsilon(i, k), \\ b(i, k) &\sim N(0, D), \\ \epsilon(i, k) &\sim N(0, \Sigma(i, k)). \end{aligned} \quad (3.5)$$

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

$$W(i, k, u) := \log(Y(i, k, u)) = x(i, k, u)\beta + z(i, k, u)b(i, k) + \epsilon(i, k, u). \quad (3.6)$$

In equation (3.6), $x(i, k, u)$ and $z(i, k, u)$ 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 model 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 equation (3.6) (and afterwards $Y(i, k, u)$), the likelihood approach starts from the BLUP for $W^*(i, k, u) := E[W(i, k, u)|b(i, k)]$. From Section 2 we know that this is given by

$$\hat{W}^*(i, k, u) = x(i, k, u)\hat{\beta} + z(i, k, u)\hat{b}(i, k), \quad (3.7)$$

with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}(i, k)$ similar to equations (2.4) and (2.5), but adjusted for the specific setup of the extensive data set on individual claims. $\hat{W}^*(i, k, u)$ is an unbiased predictor for both $W^*(i, k, u)$ and $W(i, k, u)$, in the sense that the expectations of these random variables are equal.

Following Frees, Young, and Luo (1999) the Mean Squared Error of Prediction (MSEP) is given by

$$\begin{aligned} E[\hat{W}^*(i, k, u) - W^*(i, k, u)]^2 &= \text{Var}[\hat{W}^*(i, k, u) - W^*(i, k, u)] \\ &= (\boldsymbol{x}(i, k, u)' - \boldsymbol{z}(i, k, u)' \boldsymbol{DZ}(i, k) \boldsymbol{V}(i, k)^{-1} \boldsymbol{X}(i, k)) \\ &\quad \times \left(\sum_h \boldsymbol{X}_h \boldsymbol{V}_h^{-1} \boldsymbol{X}_h' \right)^{-1} \times (\boldsymbol{x}(i, k, u)' - \boldsymbol{z}(i, k, u)' \boldsymbol{DZ}(i, k) \boldsymbol{V}(i, k)^{-1} \boldsymbol{X}(i, k))' \\ &\quad - \boldsymbol{z}(i, k, u)' \boldsymbol{DZ}(i, k) \boldsymbol{V}(i, k)^{-1} \boldsymbol{Z}(i, k) \boldsymbol{Dz}(i, k, u) + \boldsymbol{z}(i, k, u)' \boldsymbol{Dz}(i, k, u), \end{aligned} \quad (3.8)$$

where the index h runs over all claims in the data set. In the case of independent residual terms (and thus $\boldsymbol{\Sigma}(i, k)$ diagonal), an analogous expression for the Mean Squared Error of Prediction $E[\hat{W}^*(i, k, u) - W(i, k, u)]^2 = \text{Var}[\hat{W}^*(i, k, u) - W(i, k, u)]$ is obtained by replacing the last term $\boldsymbol{z}(i, k, u)' \boldsymbol{DZ}(i, k, u)$ in equation (3.8) with $\boldsymbol{z}(i, k, u)' \boldsymbol{Dz}(i, k, u) + \text{Var}(\boldsymbol{\epsilon}(i, k, u))$. 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, say, 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:

$$\begin{aligned} \hat{Y}_{GeoM}(i, k, u) &= \exp\{\hat{W}^*(i, k, u)\}, \\ \hat{Y}_{Mean}(i, k, u) &= \exp\left\{\hat{W}^*(i, k, u) + \frac{1}{2} \text{Var}(W(i, k, u) - \hat{W}^*(i, k, u))\right\}, \\ \widehat{\text{Var}}(Y(i, k, u)) &= \hat{Y}_{Mean}^2(i, k, u) [\exp\{\text{Var}(W(i, k, u) - \hat{W}^*(i, k, u))\} - 1], \end{aligned} \quad (3.9)$$

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

$$\sum_u \exp\left\{\hat{W}^*(i, k, u) + \frac{1}{2} \text{Var}(\hat{W}^*(i, k, u) - W(i, k, u))\right\}. \quad (3.10)$$

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:

$$\begin{aligned} &\sum_u \hat{Y}_{Mean}^2(i, k, u) [\exp\{\text{Var}(W(i, k, u) - \hat{W}^*(i, k, u))\} - 1] \\ &+ \sum_u \sum_{u' \neq u} \hat{Y}_{Mean}(i, k, u) \hat{Y}_{Mean}(i, k, u') \\ &\quad \times [\exp\{\text{Cov}[W(i, k, u) - \hat{W}^*(i, k, u), W(i, k, u') - \hat{W}^*(i, k, u')]\} - 1], \end{aligned} \quad (3.11)$$

where $\text{Cov}[W(i, k, u) - \hat{W}^*(i, k, u), W(i, k, u') - \hat{W}^*(i, k, u')]$ is computed in Appendix A. Expressions (3.10) and (3.11) 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 and Verrall (2002, Section 7.7) to the framework of lognormal mixed models.

3.2 Bayesian Approach: Toward a Full Predictive Distribution

Note that formulas (3.9), (3.10), and (3.11) within the likelihood context require some programming using a statistical software package (or spreadsheet) and possibly are subject to 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 provides only 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—among other papers—in de Alba (2002), Ntzoufras and Dellaportas (2002), and England and 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:

$$\begin{aligned} W(i, k)|b(i, k) &\sim N(X(i, k)\boldsymbol{\beta} + Z(i, k)b(i, k), \boldsymbol{\Sigma}(i, k)), \\ b(i, k) &\sim N(0, D), \\ \boldsymbol{\beta} &\sim N(0, F), \\ D &\sim \text{Inv-Wishart}_\nu(\mathbf{B}), \end{aligned} \quad (3.12)$$

where F is a diagonal matrix with large positive entries and \mathbf{B} is a matrix with the same dimensions as D . In the example in Section 4, we used $\nu = 3$ and $\boldsymbol{\Sigma}(i, k) = \text{diag}(\sigma_\varepsilon^2)$ with prior $\sigma_\varepsilon^2 \sim \text{Inv-Gamma}(a, b)$ and $a = b = 0.01$.

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

$$[\boldsymbol{\beta}|.] \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$$

where

$$\begin{aligned} \boldsymbol{\mu}_\beta &= \left(\sum_{i,k} X(i, k)' \boldsymbol{\Sigma}(i, k)^{-1} X(i, k) + F^{-1} \right)^{-1} \\ &\quad \times \sum_{i,k} X(i, k)' \boldsymbol{\Sigma}(i, k)^{-1} (W(i, k) - Z(i, k)b(i, k)) \end{aligned}$$

and

$$\boldsymbol{\Sigma}_\beta = \left(\sum_{i,k} X(i, k)' \boldsymbol{\Sigma}(i, k)^{-1} X(i, k) + F^{-1} \right)^{-1}, \quad (3.13)$$

next to

$$[b(i, k)|.] \sim N(\boldsymbol{\mu}_{b(i,k)}, \boldsymbol{\Sigma}_{b(i,k)}),$$

where

$$\boldsymbol{\mu}_{b(i,k)} = (Z(i, k)' \boldsymbol{\Sigma}(i, k)^{-1} Z(i, k) + D^{-1})^{-1} Z(i, k)' \boldsymbol{\Sigma}(i, k)^{-1} (W(i, k) - X(i, k)\boldsymbol{\beta})$$

and

$$\boldsymbol{\Sigma}_{b(i,k)} = (Z(i, k)' \boldsymbol{\Sigma}_{b(i,k)}^{-1} Z(i, k) + D^{-1})^{-1}, \quad (3.14)$$

and with N the total number of claims in the data set,

$$[D|.] \sim \text{Inv-Wishart}_{\nu+N} \left(\mathbf{B} + \sum_{i,k} b(i, k)b(i, k)' \right), \quad (3.15)$$

$$[\sigma_\varepsilon^2|.] \sim \text{Inv-Gamma} \left(a + \frac{1}{2} \sum_{i,k} n_{ik}, b + \frac{1}{2} \sum_{i,k} \mathbf{r}(i, k)' \mathbf{r}(i, k) \right), \quad (3.16)$$

where $r(i, k) = W(i, k) - X(i, k)\beta - Z(i, k)b(i, k)$. To perform the simulations, 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. This data set concerns Belgian automotive motor third-party liability claims. The companies involved provided those claims that were expected to be in excess of 124,000 euros, along traditional reserving techniques presently used in those companies. We first present the characteristics of the data at hand. A mixed model related to the lognormal regression model in equation (3.4) is fitted to the logarithm of the cumulative 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 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 eight 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 arrival years, 1 corresponds with 1986 and 8 with 1993. Because of 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 (3.2), with chain-ladder type structure for the mean, are considered as a benchmark for the results on the level of aggregate data. Model (3.2) was fitted to an aggregate triangle consisting of incremental payments.

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 the ‘‘arrival year’’ and DY for the ‘‘development year’’ of the claim). To avoid problems with zero or negative payments in a lognormal mixed model for individual data, our analysis models cumulative payments for individual claims. Figure 2 shows boxplots of the

Table 2

Summary of Considered Data as Classical Run-off Triangle with Cumulative Payments

Arrival Year	Development 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

Note: Lower triangle in bold: these data have to be predicted.

Figure 1

Incremental (Left) and Cumulative (Right) Payment Profiles over Development Period: Randomly Selected Claims from Data Set Summarized in Table 2 (Upper Triangle)

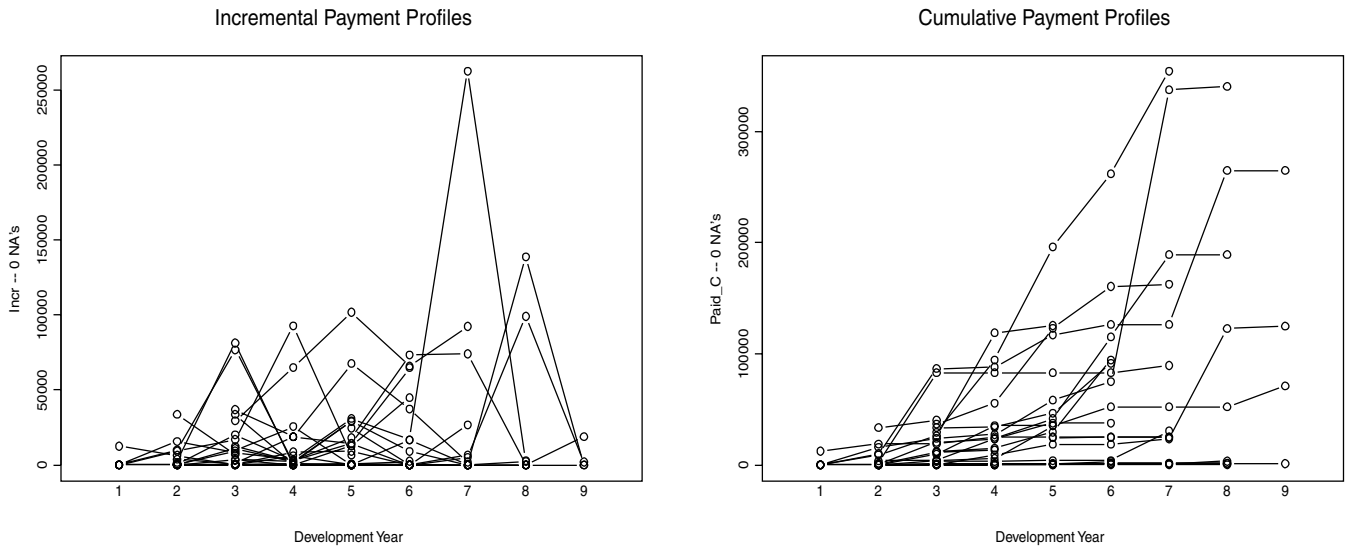
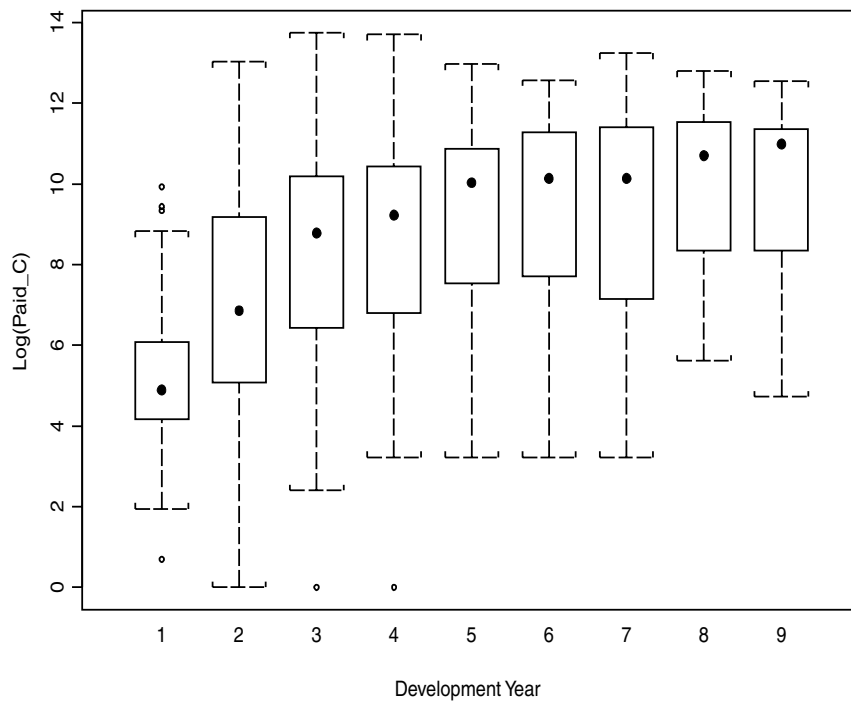


Figure 2

Boxplots of Logarithm of Cumulative Payments per Development Year, Individual Data Underlying Upper Triangle in Table 2



Note: Solid dot indicates median.

logarithmic transformed cumulative data over the available development years, for the data underlying the upper triangle in Table 2.

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 equation (3.2), are given in Table 3. These results are based on aggregate, incremental data and were obtained with WINBUGS. Within this Bayesian framework, prior specifications for the regression parameters and the variance component are similar to those discussed in Section 3.2. In this way, the values shown in the column “Mean” in Table 3 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 do not lead directly to predictions for an individual payment profile. However, practitioners often use the development factors from the deterministic chain ladder to predict individual cumulative 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 actually observed cumulative payments, on the logarithmic and the original scales, respectively. Plots 4.4 and 4.5 are on the original scale and compare the chain-ladder predictions with the mixed-model predictions as obtained with the second (Mean) and the first (Median) formulas in equations (3.9). 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 4 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 $AY + DY - 2$, where AY stands for “arrival year” and DY for “development year.” The

Table 3
**Bayesian Predictions Based on Lognormal Model with Chain-Ladder Structure for Mean
 (Data Displayed in Thousands)**

Cell	Mean	Std Dev.	MC Error	Median	Percentage Bayes. Std Err.	Percentage RMSEP	Cell	Mean	Std Dev.	MC Error	Median	Percentage Bayes. Std Err.	Percentage RMSEP
(2,9)	2,891	1,097	3.0	2,674	37.95%	44.20%	(6,9)	5,281	4,062	13.0	4,253	76.92%	124.82%
(3,8)	3,476	1,487	4.0	3,116	42.79	44.44	(7,4)	2,251	2,141	6.0	1,694	95.12	167.54
(3,9)	4,054	2,169	6.0	3,535	53.49	58.62	(7,5)	4,825	5,253	15.0	3,487	108.86	308.61
(4,7)	2,871	1,530	4.0	2,558	53.30	54.21	(7,6)	6,509	6,450	20.0	4,856	99.09	362.02
(4,8)	3,402	1,872	5.0	2,983	55.02	55.68	(7,7)	8,833	8,430	28.0	6,662	95.43	239.74
(4,9)	3,807	3,889	11.0	3,296	102.14	108.36	(7,8)	10,562	9,930	34.0	8,034	94.02	241.55
(5,6)	2,741	960	3.0	2,491	35.01	44.98	(7,9)	11,858	11,213	40.0	9,034	94.56	247.88
(5,7)	3,555	2,009	6.0	3,097	56.50	49.43	(8,3)	3,238	5,430	15.0	1,938	167.71	624.14
(5,8)	4,157	2,455	7.0	3,584	59.07	51.05	(8,4)	5,066	6,893	19.0	3,322	136.07	467.73
(5,9)	4,605	2,833	8.0	3,940	61.51	52.39	(8,5)	8,697	10,655	30.0	5,885	122.52	632.56
(6,5)	2,312	1,722	5.0	1,847	74.47	107.51	(8,6)	11,105	12,903	39.0	7,705	116.20	535.06
(6,6)	3,027	2,273	7.0	2,440	75.07	96.90	(8,7)	14,397	16,354	51.0	10,122	113.59	654.71
(6,7)	4,012	3,121	9.0	3,225	77.79	113.86	(8,8)	16,840	18,646	61.0	11,959	110.73	626.06
(6,8)	4,737	3,591	11	3,823	75.81	119.26	(8,9)	18,675	20,874	68	13,268	111.77	618.21

Notes: 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. “Percentage Bays. Std Err.” is the ratio of “Std Dev.” and “Mean.” “Percentage 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 predictive value for the cell and its real observed value, as given in Table 2.

Table 4
Mixed-Model Specification

Fixed Effects	Random Effects
$\alpha_1, \dots, \alpha_8$ Calyear DY, log (DY)	Intercept DY log (DY)

fixed effects' structure is inspired by equation (3.4) 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 $X(i, k)$ 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 calendar year, development year, and the logarithm of development year in the remaining columns. The design matrix $Z(i, k)$ 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 development year and the logarithm of development year in the remaining columns. In this way, claim-specific intercepts and slopes for development year and the logarithm of the development year are modeled.

The covariance matrix D of the random effects is not forced to satisfy any structural assumptions. The residual terms are modeled independently; thus $\Sigma(i, k)$ is diagonal. Since we are dealing with cumulative data, an AR(1) structure for $\Sigma(i, k)$ could be suggested, where AR stands for AutoRegressive. 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 and Molenberghs (2002). Moreover, we obtained better predictions with the diagonal residual matrix.

The suggested mixed model is implemented both in a likelihood-based and in 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, 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, second, the modeling of the settlement of a claim is included in the WINBUGS analysis. This is done by introducing a 0/1 indicator variable for every outstanding payment

$$Z(i, k, u) \sim \text{Bern}(1 - p(u)),$$

$$Z(i, k, u) = \begin{cases} 0 & \text{when last payment done in development year } u, \\ 1 & \text{otherwise.} \end{cases} \quad (4.1)$$

Appropriate multiplication of these indicator variables with the simulations from the posterior predictive distribution of the cumulative payments allows us to model the settlement of a claim. Here $p(j)$ ($j = 1, \dots, 9$) is the probability that a claim settles in development year j and is estimated by its empirical analogue, based on the complete data set. The estimated probabilities are given in Table 5.

Table 6 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 com-

Table 5
Estimated Probabilities of Settlement per Development Year (DY) j

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
Parameter Estimates as Obtained with PROC MIXED Analysis of Extensive Data Set Underlying Upper Triangle in Table 2

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)
	γ	-5.5011 (0.3362)
Calyear	β_1	5.1159 (0.3609)
Time	β_2	3.7983 (0.3396)
Log(Time)		
Covariance of random effects:	d_1	6.8411 (1.0015)
	d_2	0.3695 (0.1422)
	d_3	10.8311 (2.6733)
	$d_{12} = d_{21}$	0.2317 (0.2947)
	$d_{13} = d_{31}$	-3.5705 (1.3812)
	$d_{23} = d_{32}$	-1.8514 (0.5902)
Residual variance:		
Var(ϵ)	σ_ϵ^2	0.7315 (0.0492)
-2 REML log-likelihood		4260.7
AIC		4274.7

puted using the second formula in equations (3.9). The fitted values for the upper triangle are obtained as $\exp(\hat{W}^*(i, k, j) + \frac{1}{2}\text{Var}(\hat{W}^*(i, k, j) - W^*(i, k, j)))$, where j refers to an observed payment. They are displayed in Table 7 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 7 (settlement a priori) were obtained. Figure 3 illustrates the fitted profiles and predicted remaining payments for six randomly selected claims from the data set. The profiles are plotted on the logarithmic scale, together with plus/minus one standard error.

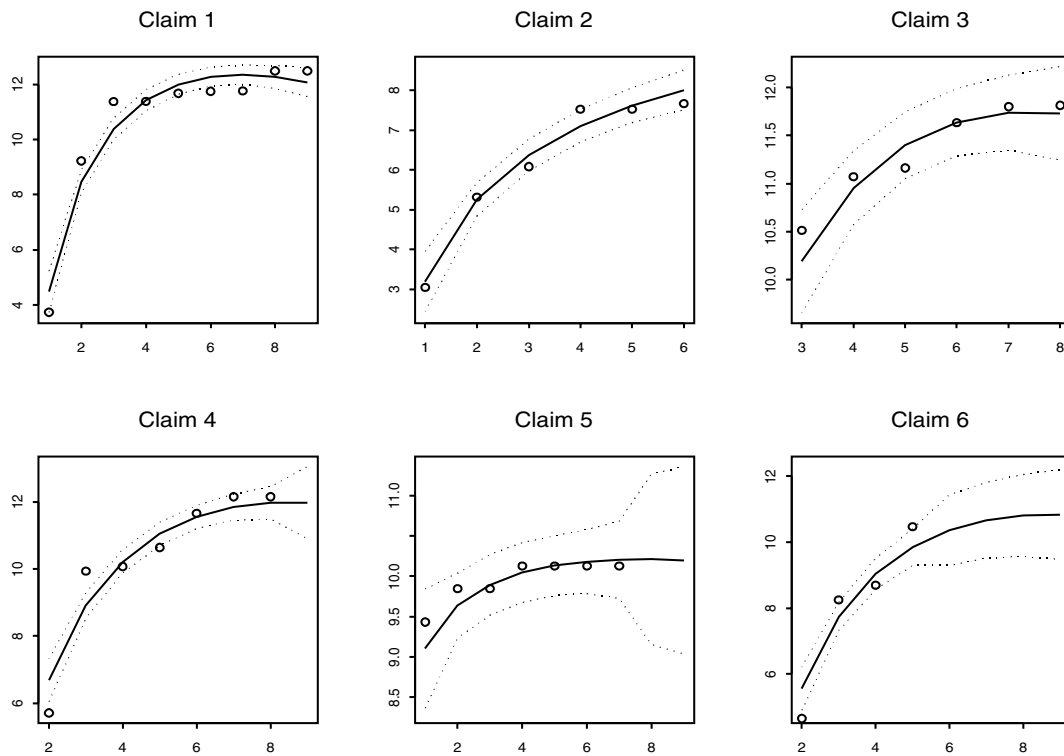
Using a Bayesian analysis, simulated values are obtained for the full predictive distribution of an individual claim, a complete arrival year, or the total reserve. Hierarchical centering and mean centering of the covariates are used, together with the prior specifications from Section 3.2. Table 8 (settlement a priori) and Table 9 (settlement modeled) 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

Table 7
Fitted Values for Upper and Predictions for Lower Triangle as Obtained with Lognormal Mixed-Model Analysis of the Extensive Data Set

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

Note: Year of settlement taken as a priori information.

Figure 3
Fitted and Predicted Payment Profiles on Log Scale, Together with ± 1 Standard Error



Notes: Circles are used for observed payments. Solid lines give fitted and predicted profiles. Dotted lines give fits/predictions ± 1 standard error as computed with PROC MIXED.

Table 8
Bayesian Predictions Obtained from Lognormal Mixed Model for Extensive Data Set Underlying Table 2 (Data Displayed in Thousands)

Cell	Mean	Std Dev.	MC Error	Median	Percentage Bayes. Std Err.	Percentage RMSEP	Cell	Mean	Std Dev.	MC Error	Median	Percentage Bayes. Std Err.	Percentage RMSEP
(2,9)	3,721	1,878	12.0	3,277	50.46	85.18	(6,9)	5,526	5,814	40.0	4,327	105.21	174
(3,8)	3,435	1,881	12.0	2,991	54.77	56.09	(7,4)	2,725	1,288	7.0	2,443	47.26	134.97
(3,9)	3,671	2,182	14.0	3,141	59.42	58.39	(7,5)	4,930	2,923	20.0	4,234	59.29	215.30
(4,7)	3,646	1,252	8.0	3,384	34.33	53.01	(7,6)	7,096	5,012	33.0	5,827	70.64	327.46
(4,8)	4,174	1,906	13.0	3,772	45.65	61.60	(7,7)	9,415	18,147	107.0	7,247	192.74	467.97
(4,9)	4,803	3,237	20.0	4,110	67.39	96.13	(7,8)	11,870	15,455	95.0	8,416	130.20	357.32
(5,6)	3,457	1,404	8.0	3,188	40.61	39.10	(7,9)	14,575	29,918	189.0	9,277	205.27	596.85
(5,7)	4,244	1,944	13.0	3,838	45.81	41.76	(8,3)	1,999	2,048	11.0	1,447	102.41	243.93
(5,8)	4,990	2,821	17	4,329	56.54	53.28	(8,4)	3,563	4,661	27	2,425	130.81	306.34
(5,9)	5,889	5,529	37.0	4,652	93.89	87.64	(8,5)	4,621	6,662	42.0	3,128	144.17	359.88
(6,5)	2,923	1,276	8.0	2,641	43.66	104.01	(8,6)	5,404	8,405	50	3,556	155.52	306.84
(6,6)	3,665	1,742	11.0	3,275	47.51	88.11	(8,7)	8,008	18,260	115	4,484	228.01	622.26
(6,7)	4,255	2,349	15	3,714	55.22	93.15	(8,8)	8,546	24,488	166	4,631	286.53	684.49
(6,8)	4,798	3,391	24.0	4,027	70.66	114.35	(8,9)	9,350	28,477	179	4,937	304.56	703.34

Notes: 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. "Percentage Bays. Std Err." is the ratio of "Std Dev." and "Mean." "Percentage 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 predictive value for the cell and its real observed value, as given in Table 2.

Table 9

Bayesian Predictions as Obtained with Lognormal Mixed Model for Extensive Data Set Underlying Table 2 (Data Displayed in Thousands)

Cell	Mean	Std Dev.	MC Error	Median	Percentage Bayes. Std Err.	Percentage RMSEP	Cell	Mean	Std Dev.	MC Error	Median	Percentage Bayes. Std Err.	Percentage RMSEP
(2,9)	3,708	1,865	12.0	3,270	50.31	84.51	(6,9)	5,602	7,831	55.0	4,033	139.78	228.71
(3,8)	3,425	1,819	12.0	2,984	53.11	54.22	(7,4)	2,737	1,351	9.0	2,443	49.36	138.76
(3,9)	3,581	2,178	16.0	3,040	60.83	58.43	(7,5)	4,381	2,630	17	3,772	60.04	184.73
(4,7)	3,647	1,233	8.0	3,403	33.80	52.48	(7,6)	6,284	4,929	31.0	5,118	78.44	299.20
(4,8)	4,005	1,660	11	3,644	41.44	52.93	(7,7)	8,217	7,728	52.0	6,294	94.06	217.18
(4,9)	4,539	3,004	20	3,902	66.19	87.62	(7,8)	10,482	14,717	102	7,338	140.40	331.66
(5,6)	3,447	1,278	8.0	3,187	37.07	37.09	(7,9)	13,409	31,011	196.0	8,022	231.26	610.73
(5,7)	4,023	1,717	12.0	3,647	42.67	39.44	(8,3)	2,009	2,206	13.0	1,464	109.80	259.30
(5,8)	4,785	2,805	17.0	4,133	58.62	53.58	(8,4)	4,394	9,237	55.0	2,680	210.21	585.73
(5,9)	5,739	5,181	35.0	4,505	90.28	82.40	(8,5)	7,178	16,764	113.0	3,935	233.53	881.67
(6,5)	2,848	1,161	7.0	2,598	40.77	96.10	(8,6)	9,216	22,450	136.0	4,805	243.59	815.32
(6,6)	3,467	1,759	11.0	3,075	50.75	84.25	(8,7)	10,356	26,556	182	5,219	256.44	905.77
(6,7)	4,018	2,485	16.0	3,433	61.84	93.63	(8,8)	11,277	32,632	197.0	5,480	289.36	918.54
(6,8)	4,681	4,928	31.0	3,731	105.29	157.48	(8,9)	12,782	133,211	804	5,701	1042.18	3242.80

Notes: 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 modeled explicitly. "Percentage Bays. Std Err." is the ratio of "Std Dev." and "Mean." "Percentage 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 predictive value for the cell and its real observed value, as given in Table 2.

triangle is summarized by its mean, median, standard deviation, and MC error. Recall that the MC error is the Monte Carlo standard error estimate of the predictive mean reserve (for a discussion, see Scollnik 2004). Other statistics, such as percentiles, can be obtained easily from WINBUGS.

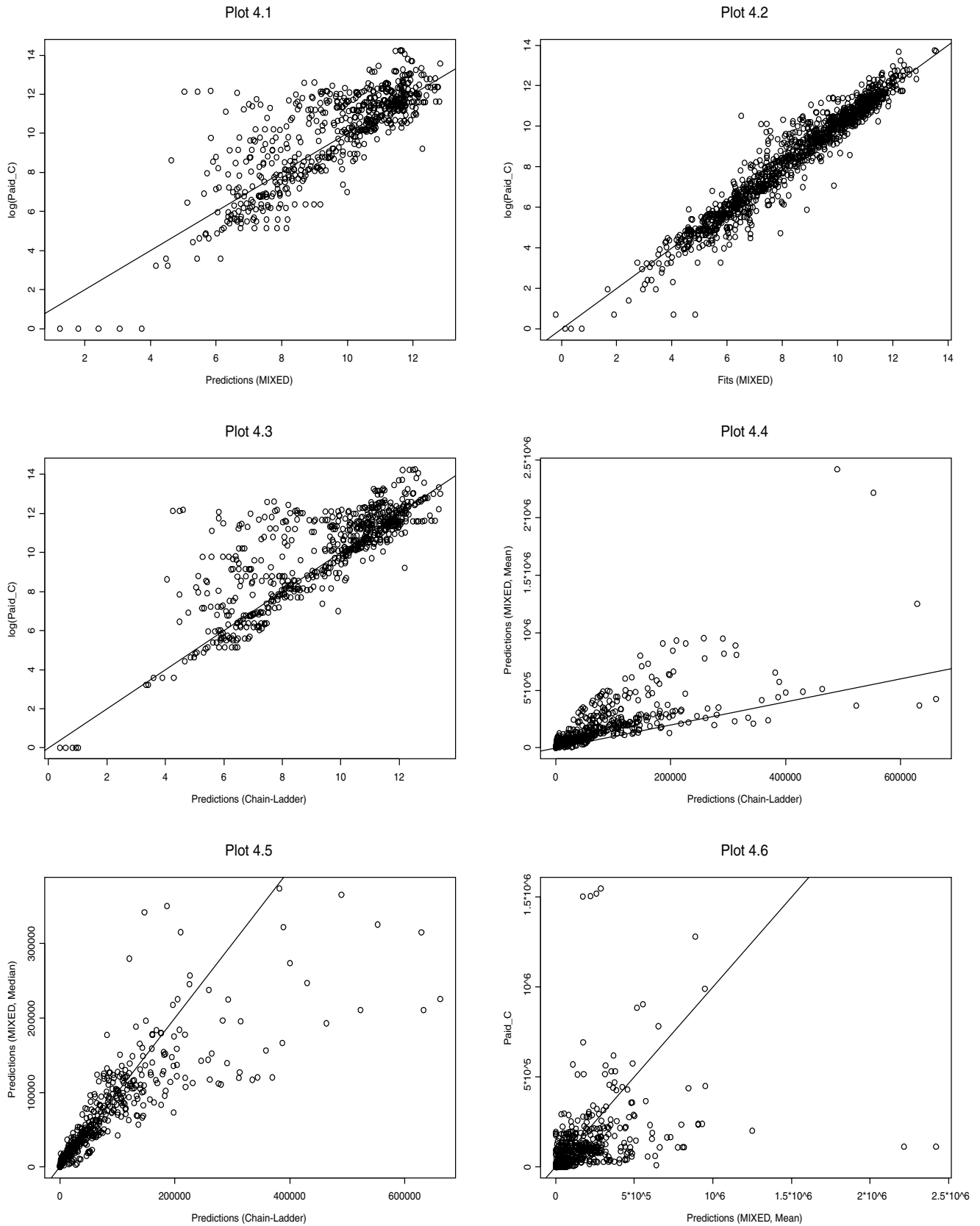
4.3 Discussion

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 really observed data. These plots illustrate that both techniques lead to comparable results on the logarithmic 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 that offers, for example, an estimate of the variability of the predictions. Furthermore, the success of the deterministic chain ladder in part can be explained by the fact that it directly uses the last observed cumulative payment in every profile.

Next, we can ask what the predictive power of the mixed model is on the level of aggregate reserves. Tables 7, 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 "Percentage Bayes. Std Err." and "Percentage RMSEP" in Table 3 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 (like (7,8), (7,9), and (8,9)), a very large standard deviation of the predictive distribution is observed.

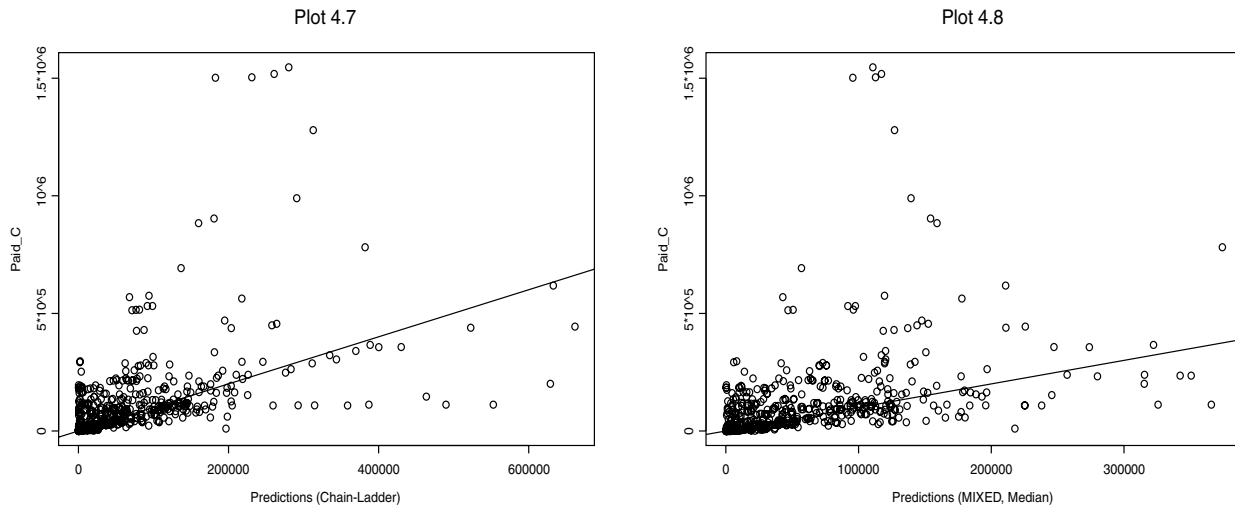
Note again that the predictions based on the mean of the lognormal distribution often severely overestimate the really observed payments, whereas use of the median leads to more reasonable predictions. This is clearly illustrated by plots 4.6 and 4.8 of Figure 4. The predictions obtained with the first formula in equations (3.9) also compare favorably with respect to the chain-ladder predictions, in contrast to those obtained with the second formula in equations (3.9); see plots 4.4–4.8. 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 eqs. (3.9))—is a general problem in reserving models within the lognormal framework; see, for example, England and Verrall (2002).

Figure 4
Comparison between Fitted and Predicted Payments and Real Observations



Notes: Plots 4.1, 4.2, and 4.3 are on the log scale. Plots 4.4, 4.5, 4.6, 4.7, and 4.8 are on the original scale of the payments. A line with 0 intercept and slope 1 is added in each plot.

Figure 4
(continued)



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 and 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 modeling of zeros. A stochastic discounting process also can be included.

APPENDIX

COVARIANCE EXPRESSION IN EXPRESSION (3.11)

In this Appendix we describe how an expression for the covariance in expression (3.11) can be derived. In the sequel it is assumed that the residual terms are modeled independently (with $\Sigma(i, k) = \text{diag}(\sigma_{\epsilon}^2)$), but the formulas can be generalized in a straightforward way:

$$\begin{aligned} & \text{Cov}[W(i, k, u) - \hat{W}^*(i, k, u), W(i, k, u') - \hat{W}^*(i, k, u')] \\ &= \text{Cov}[W(i, k, u), W(i, k, u')] - \text{Cov}[W(i, k, u), \hat{W}^*(i, k, u')] \\ & \quad - \text{Cov}[\hat{W}^*(i, k, u), W(i, k, u')] + \text{Cov}[\hat{W}^*(i, k, u), \hat{W}^*(i, k, u')]. \end{aligned} \tag{A.1}$$

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

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

Some matrix calculations lead to

$$\begin{aligned} \text{Cov}[W(i, k, u), \hat{W}^*(i, k, u')] &= \mathbf{z}(i, k, u)' \mathbf{DZ}(i, k)' V(i, k)^{-1} X(i, k) \left(\sum_h X_h' V_h^{-1} X_h \right)^{-1} \\ &\quad \times \{ \mathbf{x}(i, k, u')' - \mathbf{z}(i, k, u')' \mathbf{DZ}(i, k)' V(i, k)^{-1} X(i, k) \}' \\ &\quad + \mathbf{z}(i, k, u)' \mathbf{DZ}(i, k)' V(i, k)^{-1} \mathbf{Z}(i, k) \mathbf{Dz}(i, k, u'), \end{aligned} \quad (\text{A.3})$$

and

$$\begin{aligned} \text{Cov}[\hat{W}^*(i, k, u), W(i, k, u')] &= \{ \mathbf{x}(i, k, u)' - \mathbf{z}(i, k, u)' \mathbf{DZ}(i, k)' V(i, k)^{-1} X(i, k) \} \left(\sum_h X_h' V_h^{-1} X_h \right)^{-1} \\ &\quad \times X(i, k)' V(i, k)^{-1} \mathbf{Z}(i, k) \mathbf{Dz}(i, k, u') \\ &\quad + \mathbf{z}(i, k, u)' \mathbf{DZ}(i, k)' V(i, k)^{-1} \mathbf{Z}(i, k) \mathbf{Dz}(i, k, u'), \end{aligned} \quad (\text{A.4})$$

Furthermore

$$\begin{aligned} \text{Cov}[\hat{W}^*(i, k, u), \hat{W}^*(i, k, u')] &= \text{Cov}[\mathbf{x}(i, k, u)' \hat{\boldsymbol{\beta}}, \mathbf{x}(i, k, u')' \hat{\boldsymbol{\beta}}] + \text{Cov}[\mathbf{z}(i, k, u)' \hat{\mathbf{b}}(i, k), \mathbf{z}(i, k, u')' \hat{\mathbf{b}}(i, k)], \end{aligned}$$

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

$$\begin{aligned} \text{Var}[\hat{\mathbf{b}}(i, k)] &= \mathbf{DZ}(i, k)' V(i, k)^{-1} \left\{ V(i, k) - X(i, k) \left(\sum_h X_h' V_h^{-1} X_h \right)^{-1} X(i, k)' \right\} V(i, k)^{-1} \mathbf{Z}(i, k) \mathbf{D}, \end{aligned}$$

and

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_h X_h' V_h^{-1} X_h \right)^{-1},$$

from which we conclude

$$\begin{aligned} \text{Cov}[\hat{W}^*(i, k, u), \hat{W}^*(i, k, u')] &= \mathbf{x}(i, k, u)' \left(\sum_h X_h' V_h^{-1} X_h \right)^{-1} \mathbf{x}(i, k, u') + \mathbf{z}(i, k, u)' \mathbf{DZ}(i, k)' V(i, k)^{-1} \\ &\quad \times \left\{ V(i, k) - X(i, k) \left(\sum_h X_h' V_h^{-1} X_h \right)^{-1} X(i, k)' \right\} V(i, k)^{-1} \mathbf{Z}(i, k) \mathbf{Dz}(i, k, u'). \end{aligned} \quad (\text{A.5})$$

Combining equations (A.2), (A.3), (A.4), and (A.5) into equation (A.1) then leads to an expression for the covariance term in expression (3.11).

ACKNOWLEDGMENTS

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

REFERENCES

- BARNETT, GLEN, AND BEN ZEHNWIRTH. 1998. Best Estimates for Reserves. *Casualty Actuarial Science Forum* 1–54.
 DE ALBA, ENRIQUE. 2002. Bayesian Estimation of Outstanding Claims Reserves. *North American Actuarial Journal* 6(4): 1–20.

- DEMIDENKO, EUGENE. 2004. *Mixed Models: Theory and Applications*. Hoboken, NJ: Wiley.
- DE VYLDER, FLORIAN, AND MARC J. GOOVAERTS. 1979. Proceedings of the First Meeting of the Contact Group "Actuarial Science," *KU Leuven*, nr. 7904B, *wettelijk depot*: D/1979/23761/5.
- DIGGLE, PETER J., PATRICK HEAGERTY, KUNG-YEE LIANG, AND SCOTT L. ZEGER. 2002. *Analysis of Longitudinal Data*. Oxford: Oxford University Press.
- DORAY, LOUIS G. 1996. UMVUE of the IBNR Reserve in a Lognormal Linear Regression Model. *Insurance: Mathematics and Economics* 18(1): 43–57.
- ENGLAND, PETER D., AND RICHARD J. VERRALL. 2002. Stochastic Claims Reserving in General Insurance. *British Actuarial Journal* 8: 443–544.
- FREES, EDWARD W., VIRGINIA R. YOUNG, AND YU LUO. 1999. A Longitudinal Data Analysis Interpretation of Credibility Models. *Insurance: Mathematics and Economics* 24(3): 229–47.
- . 2001. Case Studies Using Panel Data Models. *North American Actuarial Journal* 4(4): 24–42.
- HAASTRUP, SVEND, AND ELJA ARJAS. 1996. Claims Reserving in Continuous Time: A Nonparametric Bayesian Approach. *ASTIN Bulletin* 26(2): 139–64.
- KREMER, ERHARD. 1982. IBNR Claims and the Two Way Model of ANOVA. *Scandinavian Actuarial Journal* 47–55.
- KUNKLER, MICHAEL. 2004. Modelling Zeros in Stochastic Reserving Models. *Insurance: Mathematics and Economics* 34(1): 23–35.
- LAIRD, NAN M., AND JAMES H. WARE. 1982. Random-Effects Models for Longitudinal Data. *Biometrics* 38: 963–74.
- NORBERG, RAGNAR. 1993. Prediction of Outstanding Liabilities in Non-life Insurance. *ASTIN Bulletin* 23(1): 95–115.
- NTZOUFRAS, IOANNIS, AND PETROS DELLAPORTAS. 2002. Bayesian Modeling of Outstanding Liabilities Incorporating Claim Count Uncertainty. *North American Actuarial Journal* 6: 113–28.
- SCOLLNIK, DAVID P. M. 2004. Bayesian Reserving Models Inspired by Chain-Ladder Methods and Implemented Using WINBUGS. *ARCH* 2004(2).
- TAYLOR, GREG AND MIREILLE CAMPBELL. 2002. Statistical Case Estimation. Working paper. www.economics.unimelb.edu.au/actwww/html/no104.pdf.
- TAYLOR, GREG, GRAINNE MCGUIRE, AND ALAN GREENFIELD. 2003. Loss Reserving: Past, Present and Future. Working paper. www.economics.unimelb.edu.au/actwww/html/no109.pdf.
- VERBEKE, GEERT, AND GEERT MOLENBERGHS. 1997. *Linear Mixed Models in Practice: A SAS Oriented Approach*. New York: Springer.
- . 2000. *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- ZEHNWIRTH, BEN. 1985. *ICRFS Version 4 Manual and Users Guide*. Benhar Nominees Pty Ltd, Turrumurra, NSW, Australia.

Discussions on this paper can be submitted until July 1, 2006. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.