STOCHASTIC CLAIMS RESERVING IN INSURANCE USING RANDOM EFFECTS

Michal Gerthofer, Michal Pešta*

Abstract
Estimation of claims reserves, which should be held by the insurer so as to be able to meet expected future claims arising from policies currently in force and policies written in the past, presents an important task for insurance companies to predict their liabilities. A common approach to the reserving problem is based on generalized linear models (GLM). In this article, the application of generalized linear mixed models (GLMM) – an extension of the GLM – for estimation of the loss reserves is shown. Since the GLMM allows incorporating a random effect instead of several fixed effects corresponding to the accident years as in case of the GLM, volatility of the prediction is reduced. This allows more flexible risk valuation, which is a crucial element of risk management and capital allocation practices of non-life insurers. A real data example together with diagnostics for the model selection are provided as an illustration of the potential benefits of the presented approach.

Keywords: claims reserving, non-life insurance, dependency modelling, random effects, mixed models, GLM, GLMM, panel data
JEL Classification: C13, C18, C23, C33, C51, G22

1. Introduction

Claims reserving is a classical problem in non-life insurance, sometimes also called general insurance (in the UK) or property and casual insurance (in the USA). A non-life insurance policy is a contract between the insurer and the insured. The insurer receives a deterministic amount of money, known as premium, from the insured in order to obtain a financial coverage against well-specified randomly occurred events. If such an event (claim, loss) happens, the insurer is obliged to pay in respect of the claim a claim amount, also known as loss amount.

Claims reserving now means, that the insurance company puts sufficient provisions from the premium payments aside, so that it is able to settle all the claims that are caused by these insurance contracts. The main issue is how to determine or estimate these claims reserves, which should be held by the insurer. A number of various methods has been invented in this field, see England and Verrall (2002) or Wüthrich and Merz (2008) for an overview.

* Michal Gerthofer, Faculty of Informatics and Statistics, Department of Statistics and Probability, University of Economics in Prague, Prague, Czech Republic (Michal.Gerthofer@gmail.com); Michal Pešta, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Charles University in Prague, Prague, Czech Republic (Michal.Pesta@mff.cuni.cz).
The research of Michal Gerthofer was supported by Grant No. IGA c. 70/2016.
The research of Michal Pešta was supported by the Czech Science Foundation Project GAČR No. 15-04774Y.
The authors would like to thank the referees for careful reading of their manuscript and providing suggestions that improved this manuscript.

1.1 Motivation and aims

Generalized linear models (GLM) have become a common statistical tool for the whole class of actuarial methods in non-life insurance. All the classical approaches for claims reserving are based on the assumption that the impact of each accident year and each development year of a claim on the claim amount should be represented by one fixed effect. However, this assumption can be sometimes unrealistic or at least questionable. Antonio and Beirlant (2007) suggest the generalized linear mixed models (GLMM) for the claims reserving purposes, which extends the classical GLM and is frequently used in panel (longitudinal) data analyses. Nevertheless, the GLMM approach was proposed in a way that more granular data are required (*i.e.* claim-by-claim data) and several quite restrictive assumptions are needed. Here, we demonstrate for the first time how to use the GLMM for claims reserving on the aggregated (triangular) data without prior knowledge (*i.e.* expert judgement) on the premium.

Claims reserving is a worldwide phenomenon that is of high interest among people from the insurance business. The Solvency II directive brings the stochastic claims reserving problem not only inside the insurance companies (and this is done by law in many countries). The stochastic claims reserving becomes an issue for the auditors as well as for the supervisors. Such an important economic problem is more than relevant not only for the academic society, but also for many practitioners. Indeed, many of the insurance oriented economic experts can benefit from the proposed methods: practitioners, who can immediately utilize a new promising approach in stochastic claims reserving process; and theoreticians, who may develop and, consequently, prove the required statistical inference within the proposed framework.

1.2 Structure of the paper

The principles of the GLMM as a modelling technique for panel data together with corresponding parameter estimation methods are explained in Section 2. In Section 3, the claims reserving problem for triangular data is introduced. Section 4 elaborates the core of this paper – possible applications of the GLMM within the claims reserving setup. A real data example is presented in Section 5 in order to illustrate potential benefits of incorporating the random effects into the stochastic claims reserving models.

2. Panel Data Framework

Panel data contain observations of multiple instances obtained over multiple time periods for the same individuals or subjects. Time series and cross-sectional data are special cases of panel data. Cross-sectional data are observations of different individuals or subjects on the same occasion. On the other hand, time series capture the development of one individual during a time period.

In general, cross-sectional data does not have a special order and is commonly considered independent, which is in contrast with chronologically ordered time series. It is also assumed that the structure of panel data is the same during collecting and individuals are similar in a certain way. Due to this, we can apply a model with the same structure on all of them.

2.1 Linear mixed models

Suitable and often used statistical models for panel data are Linear Mixed Models (LMM), which allow us to handle modelling both *fixed effects and random effects*. Mathematical formulation is given by

$$Y_{it} = \boldsymbol{X}_{it}^T \boldsymbol{\beta} + \boldsymbol{Z}_{it}^T \boldsymbol{b}_i + \varepsilon_{it}$$

or equivalently in the vector-matrix[1] notation

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

where $t = 1, \ldots, n_i$ and $i = 1, \ldots, N$. Here, the different number of observations through subjects i's is allowed, corresponding to the general unbalanced design. Moreover, $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ is a vector of the response for the i-th individual (the outcome variable of interest), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of the common fixed-effects (regression coefficients), $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in_i})^T$ is a fixed regression matrix of the predictor variables (regressors), $\boldsymbol{X}_{it} = (X_{1it}, \ldots, X_{pit})^T$, $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})^T$, is a vector of the random effects (the random complement to the fixed $\boldsymbol{\beta}$), $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \ldots, \boldsymbol{Z}_{in_i})^T$ is a design matrix for the random effects, and $\boldsymbol{Z}_{it} = (Z_{1it}, \ldots, Z_{qit})^T$. Random vector $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, \ldots, \boldsymbol{\varepsilon}_{in_i})^T$ is independent with vector $\boldsymbol{b}_i$ and both have multivariate normal distribution with zero mean and variance matrices $\boldsymbol{R}_i$ and $\boldsymbol{G}$, respectively. Joint vectors $(\varepsilon_i^{T}, \boldsymbol{b}_i^T)^T$ are independent with respect to i. Additionally, we assume $\boldsymbol{R}_i = \sigma_\varepsilon^2 \boldsymbol{I}_{n_i}$, because it is not possible to simultaneously estimate unstructured $\boldsymbol{G}$ and $\boldsymbol{R}_i$.

Assume covariance matrix of dependent variable is known and has the following form

$$Cov(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i = Cov(\boldsymbol{Z}_i \boldsymbol{b}_i) + Cov(\boldsymbol{\varepsilon}_i) = \boldsymbol{Z}_i \boldsymbol{G} \boldsymbol{Z}_i^T + \boldsymbol{R}_i$$

Then unknown regression parameters $\boldsymbol{\beta}$ can be estimated by applying Generalized Least Squares (GLS), which leads to the following estimate

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Y}_i \right).$$

This is Best Linear Unbiased Estimate (BLUE) of $\boldsymbol{\beta}$. For more details and properties of the estimate, see *e.g.* Rao *et al*. (1999), Chapter 4.

In cases where we do not know $\boldsymbol{\Sigma}_i$, we have to find a consistent estimate $\hat{\boldsymbol{\Sigma}}_i$ using Maximum Likelihood (ML) or Restricted ML (REML). Then, this estimate can be used for the estimation of $\boldsymbol{\beta}$ in Feasible Generalized Least Squares (FGLS)

$$\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{\Sigma}}_l^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \hat{\boldsymbol{\Sigma}}_l^{-1} \boldsymbol{Y}_i \right).$$

1 Note that the following notation $Y_{it} \equiv Y_{i,t}$ and $Y_{itj} \equiv Y_{i,t,j}$ is used for indices. Thus, where two indices stay side by side, a comma should be imagined and there is no multiplication of indices. For special cases where sum occurs in index, the same rule is applied, *i.e.*, $Y_{1+it+n} \equiv Y_{1+i,t+n}$. The same holds for numbers in indices due to the fact that we only use integers less than 10 in indices, hence, notation is given by $Y_{12} \equiv Y_{1,2}$ or $Y_{1n-1} \equiv Y_{1,n-1}$.

See more about FGLS in Greene (2002), Chapter 10.5. Iterative Generalized Least Squares (IGLS) can be applied as well to this procedure. It is based on iterations between the GLS estimate of $\boldsymbol{\beta}$ for given estimate of covariance matrix and consequently re-estimation of $\hat{\boldsymbol{\Sigma}}_i$. This process is repeated until the required precision is obtained.

In many applications, inference is focussed on the fixed effects $\boldsymbol{\beta}$ because of their interpretation in terms of changes in the mean response over time. However, we may want to predict an individual specific response profile, *e.g.* we may want to identify those individuals who showed the greatest increase or decrease in the response over time. The structure of this model allows us to estimate (predict) an individual specific response. Prediction of random variable translates into the problem of predicting the conditional mean of $\boldsymbol{b}_i$, given the vector of response $\boldsymbol{Y}_i$. Using properties of join multivariate normal distribution, it can be written as

$$E\left(\boldsymbol{b}_i \middle| \boldsymbol{Y}_i\right) = \boldsymbol{G}\boldsymbol{Z}_i^T \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right).$$

This is known as Best Linear Unbiased Predictor (BLUP). In practice, this predictor is unusable due to unknown variance matrices as in the previous case, but they can be replaced by REML (ML) estimates. Then, we get empirical BLUP or "empirical Bayes" estimator

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{G}}\boldsymbol{Z}_i^T \hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}).$$

Predicted response profile is given by

$$\hat{\boldsymbol{Y}}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i,$$

which can be rewritten as follows

$$\hat{\boldsymbol{Y}}_i = \left(\hat{\boldsymbol{R}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}\right)\boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \left(\boldsymbol{I}_{n_i} - \hat{\boldsymbol{R}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}\right)\boldsymbol{Y}_i.$$

This expression shows how the empirical Bayes estimator "shrinks" the *i*-th subject's predicted response profile to the population-average mean response profile. If the within-subject variability $\boldsymbol{R}_i$ is large relatively to the between-subject variability $\boldsymbol{\Sigma}_i$, more weight is given to $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ than to the *i*-th observed response.

2.2 Generalized linear models

The response (dependent variable) in the LMM is normally distributed. This assumption is very restrictive for many real situations. Thus, we relax this restriction in order to handle *dependent variables from various distributions*. Generalized Linear Models (GLM) deal with responses, whose distributions belong to exponential family. *Exponential family* contains distributions with densities that can be written as

$$f\left(Y\middle|\theta,\varphi\right) = \exp\left\{\frac{Y\theta - b\left(\theta\right)}{\varphi} + c\left(Y,\varphi\right)\right\}, \tag{2}$$

where $\theta \in \mathbb{R}$ is a canonical parameter, $\varphi \in (0,\infty)$ is a dispersion parameter, and $b(\cdot)$, $c(\cdot,\cdot)$ are real functions. The stated form of distribution is called canonical. Normal, gamma, inverse Gaussian, Poison, and alternative distribution are some of the members of this family.

Assume random variable Y follows a distribution from exponential family and $b(\cdot)$ is twice continuously differentiable. Then the moment generation function of Y exists, is finite, and is equal to

$$m_Y(t) = E\exp\{tY\} = \exp\left\{\frac{b(t\varphi+\theta)+b(\theta)}{\varphi}\right\}.$$

Consequently, since $b(\cdot)$ is twice continuously differentiable, $m_y(t)$ is also twice differentiable at zero. Using property of moment generation function, we can obtain the following assertions

$$E(Y) = \mu = b'(\theta) \quad (<\infty),$$

$$Var(Y) = \varphi b''(\theta) \quad (<\infty). \tag{3}$$

Corollaries from (2) ($\varphi > 0$) and (3) ($Var(Y) > 0$) are that $b(\cdot)$ is convex function and $b'(\cdot)$ is strictly increasing. Hence, $b'(\cdot)$ has a well-defined inverse. Variance function $V(\cdot)$ can be defined, for which $Var(Y) = \varphi V(\mu)$ and $b''(\theta) = V[b'(\theta)]$.

To estimate the parameters of the exponential family distribution, let $Y_1, \ldots, Y_n$ be a random sample from the distribution with density (2). The ML method can be used for estimation of θ. In Lehmann (1983), Chapter 6.4, it is discussed that estimate is consistent and asymptotically normal if regularity conditions are satisfied. The moment method can be applied to obtain the estimate of φ as well. Moreover, the structure of the joint information matrix for the vector $(\theta,\varphi)^T$ implies asymptotic independence of ML estimates $\hat{\theta}$ and $\hat{\varphi}$, when regularity conditions for the vector $(\theta,\varphi)^T$ are satisfied.

The aim of this paper is not to describe the GLM, so the definition, maximum likelihood estimation in the GLM as well as a numerical algorithm used in order to solve the likelihood equations can be found *e.g.* in Dobson (2002).

2.3 Generalized Linear Mixed Models

In the previous, when we described the LMM and by using individual random effect, we were able to introduce the within-subject randomness. Random effects give specific *individual volatility to the panels* adding flexibility to this approach. Furthermore, the "empirical Bayes" estimator for individual random effects was listed. Nevertheless, it still has a big limitation on the distribution of the response. We also introduced the GLM, which allows to model response variable from exponential family and a more complex mean structure, but its disadvantage is that the within-subject observations of these variables are assumed to be independent. Due to these facts Generalized Linear Mixed Model (GLMM) is proposed. It *combines all the benefits from LMM and GLM*. GLMM is given by the following conditions:

- We assume that Y_{it} follows unbalanced design with N individuals and n_i measurements for each of them, like in the LMM. Furthermore, independence between individuals is assumed as well, *i.e.* $\boldsymbol{Y}_i$ is independent with $\boldsymbol{Y}_j$ for $i \neq j$.

- Next, the random effects $\boldsymbol{b}_i$, $i = \{1, ..., N\}$ are independent random vectors and $\boldsymbol{b}_i \sim \mathcal{N}_q(0, \boldsymbol{D})$, where $\boldsymbol{D} \equiv \boldsymbol{D}(\psi)$ depends on parameter ψ.
- Given $\boldsymbol{b}_i$, components of $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^T$ are conditionally independent, with density belonging to exponential family distribution

$$f\left(Y_{it} \middle| \boldsymbol{b}_i\right) = \exp\left\{\frac{Y_{it}\theta_{it} - b\left(\theta_{it}\right)}{\varphi} + c\left(Y_{it}, \varphi\right)\right\}.$$

- The conditional mean of Y_{it} given $\boldsymbol{b}_i$ is

$$\mu_{it} \equiv E(Y_{it} | \boldsymbol{b}_i) = b'(\theta_{it})$$

and the conditional variance of Y_{it} given $\boldsymbol{b}_i$ has the following form

$$Var(Y_{it} | \boldsymbol{b}_i) = \varphi b''(\theta_{it}) \equiv \varphi V(\mu_{it}).$$

- Furthermore, it is assumed that μ_{it} is related to the linear predictor

$$\eta_{it} = \boldsymbol{X}_{it}^T \boldsymbol{\beta} + \boldsymbol{Z}_{it}^T \boldsymbol{b}_i \tag{4}$$

through the link function $g(\mu_{it}) = \eta_{it}$.

Conditional Y_{it}, given $\boldsymbol{b}_i$, satisfies the GLM and the inclusion of $\boldsymbol{b}_i$ in all η_{it} brings in correlation between $Y_{i1}, Y_{i2}, ..., Y_{in_i}$ like in the LMM.

Due to the assumption of conditional distribution of dependent variables, the maximum likelihood method can be used. Unfortunately, this likelihood does not generally have a closed-form solution and approximation methods for estimation must be used. The likelihood is given by

$$L(\boldsymbol{\beta}, \psi | \boldsymbol{Y}) = \prod_{i=1}^{N} f(\boldsymbol{Y}_i | \boldsymbol{\beta}, \psi) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{t=1}^{n_i} f(\boldsymbol{Y}_{it} | \boldsymbol{\beta}, \boldsymbol{b}_i) f_b(\boldsymbol{b}_i | \psi) \mathrm{d}\boldsymbol{b}_i. \tag{5}$$

Equation 5 can be expressed using canonical link $g(\cdot)$ and multivariate normal distribution f_b as follows

$$L(\boldsymbol{\beta}, \psi | \boldsymbol{Y}) = \prod_{i=1}^{N} (2\pi)^{-q/2} |\boldsymbol{D}|^{-1/2} \int_{\mathbb{R}^q} \exp\left\{\frac{1}{\varphi}\left[\boldsymbol{Y}_i^T (\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) - 1_{n_i}{}^T b(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)\right]\right\}$$
$$\times \exp\left\{1_{n_i}{}^T k\left(\boldsymbol{Y}_i, \varphi\right) - \frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{D}^{-1} \boldsymbol{b}_i\right\} \mathrm{d}\boldsymbol{b}_i,$$

where the functions $b(\cdot)$ and $k(\cdot)$ are applied to vectors by element-wise calculation. Next, log-likelihood is in form

$$\ell(\boldsymbol{\beta}, \psi | \boldsymbol{Y}) = -\frac{N}{2} \log |\boldsymbol{D}| + \sum_{i=1}^{N} \log \int_{\mathbb{R}^q} \exp\left\{h(\boldsymbol{\beta}, \psi, \varphi, \boldsymbol{Y})\right\} \mathrm{d}\boldsymbol{b}_i + C,$$

where C is constant with respect to $\boldsymbol{\beta}$, ψ, and $h(\cdot)$ is function of $\boldsymbol{\beta}$, ψ, φ, $\boldsymbol{Y}$. There are many approaches to maximize $\ell\left(\boldsymbol{\beta},\psi\middle|\boldsymbol{Y}\right)$. Traditional one is based on the Laplace approximation with after Gaussian integration, cf. Raudenbush (2000). Other approaches can be also used, *e.g.* numerical integration techniques, Gaussian quadrature (GQ) described in McCulloch and Searle (2001), Chapter 10.3, which approximate the integral appealing in (5) as weighted sum of a specified number of quadrature points for each dimension of the integration. More quadrature points mean an increase in accuracy of approximation, however, it causes higher computational demands, where we have some limitations. Thus, an appropriate balance between accuracy and optimality must be chosen. In order to maximize such approximation, Newton-Raphson can be used. More information can be found in Rabe-Hesketh and Skrondal (2002). Marginal quasi-likelihood (MQL), Penalized Quasi-Likelihood (PQL), Markov Chain Monte Carlo (MCMC) and Adaptive Gaussian Quadrature (AGQ) are other commonly used methods for computing ML estimates. These approaches are described in Diggle *et al.* (2002), Chapter 4.6 and in McCulloch and Searle (2001), Chapter 10.3. After calculating estimates of $\boldsymbol{\beta}$, $\boldsymbol{D}$, φ, the prediction for $\boldsymbol{b}_i$ can be obtained by

$$\hat{\boldsymbol{b}}_i = E(\boldsymbol{b}_i \middle| \boldsymbol{Y}_i, \hat{\boldsymbol{\beta}}, \hat{\varphi}, \hat{\boldsymbol{D}}),$$

which coincides with the empirical "Bayes estimator" or BLUP for the LMM. Such prediction is also not easy to obtain due to integration over the distribution of the unobserved random effects, $\boldsymbol{b}_i$, and again numerical methods must be used. Prediction of $\boldsymbol{b}_i$ is heavily influenced by the normal distribution assumption of random effect. Thus, the prediction is very sensitive to misspecification of the distribution. However, this misspecification does not produce a discernible bias for estimates of the fixed effects. On the other hand, estimates of fixed effects can be severely biased when the variance of random effects depends upon the subject, cf. Fitzmaurice *et al*. (2004), Chapter 12.4.

3. Introduction to Reserving Theory

This section deals with claims reserving, which is an important task in non-life insurance. Non-life insurance offers financial coverage against various types of random occurrences in case that well-specified event happens. The value, which the insurer is obligated to pay as coverage, is called the *claim amount or the loss amount*. According to the type of claim, non-life insurance is split into several Lines of Business (LoB), *e.g.* motor/car insurance, property insurance, liability insurance, accident insurance, *etc*. Number and types of LoBs vary through different insurance companies.

Reserving in non-life insurance needs a special approach because of a *time lag* between claims occurrence and claims reporting to the insurer, which is called reporting delay. It can also take several years until the process is finally closed after the claim is reported. It is also possible that an already closed claim will need to be reopened because of new facts. Due to the mentioned time lag, the claim cannot be settled right after its accident day and the so-called claims reserves have to be created. These reserves should represent all future claims arising from policies currently in force and policies written in the past. This amount of money should be held by the insurance company with the aim to meet their future liabilities.

It is worth mentioning that claims costs are often impacted by inflation. The main effect of inflation is not related to the salary or price but to the specifications of a particular LoB. For example, in the motor hull LoB, it is driven by the complexity of car repairing techniques and in LoB accident insurance, it is driven by improvements in medical care or in medicine. The impact of inflation develops through accident years as well as development years.

3.1 Reserving terminology and notation

In this section, we introduce the classical claims reserving notations and terminology. Reserving approaches are based on history of claims. In order to capture all this information in standardized form, the so-called claims development triangle is used (Table 1). Let Y_{it} stands for all the claim amounts in development year t with accident year i. We refer to Y_{it} as incremental claims in accident year i made in the accounting year $i + t$. Then current year n corresponds to the most recent accident year as well as the most recent development year. The history of claims is placed in *right-angled isosceles triangle* $\{Y_{it}\}$, where $i = 1, \ldots, n$ and $t = 1, \ldots, n + 1 - i$.

Table 1 | Run-off Triangle for Incremental Claim Amounts Y_{it}

Accident	**Development year *t***						
Year *i*	**1**	**2**	**…**	***t***	**…**	***n* – 1**	***n***
1	Y_{11}	Y_{12}	…	Y_{1t}	…	Y_{1n-1}	Y_{in}
2	Y_{21}	Y_{22}	…	Y_{2t}	…	Y_{2n-1}	
⋮	⋮	⋮	⋱				
i	Y_{i1}	Y_{i2}		Y_{it}			
⋮	⋮	⋮					
***n* – 1**	Y_{n-11}	Y_{n-12}					
n	Y_{n1}						

Source: Author

Let us denote a random variables C_{it}, cumulative payments or cumulative claims, in origin year i after t development years, *i.e.* $C_{it} = \Sigma_{k=1}^{t} Y_{ik}$. All our effort is concentrated on estimating the ultimate claim amount C_{in} and, consequently, on calculating reserves for all accident years $i = 2, \ldots, n$ as follows

$$R_i^{(n)} = C_{in} - C_{in+1-i}. \tag{6}$$

So the main goal is to statistically predict the reserves based on the triangle of claim amounts. This text deals only with reserves defined in (6) and does not assume any tail factor.

4. Claims Reserving within the Panel Data Framework

Previous sections described all the necessary theory in general. In this chapter, theory of the GLMM is applied on the claims reserving problem. The advantages of these models seem to be suitable when more modelling flexibility within accident year i are needed and possible dependencies among the incremental claims within accident year i exist. Another very important property is that the *effect for accident year can be taken as random*. This fact allows us to predict these random variables and we do not have to spend parameters on them. In case of random intercept, only the variance of this random variable must be estimated, which means to estimate one unknown parameter unlike $n - 1$ parameters for each accident year as in the GLM.

Besides using the GLMM, Hudecová and Pešta (2013) proposed another modelling technique for panel data in the reserving framework, which is called Generalized Estimating Equations (GEE). On the one hand, the GEE do not assume a specified distribution of the outcome. On the other hand, a correlation structure is required in advance for the GEE. Therefore, their approach may be considered as distribution free one. Pešta and Okhrin (2014) also applied growth curves for the run-off triangles as another method for handling panel data.

4.1 GLMM method for claims reserving

Claims reserving using GLMM is going to be described. Focus lies on a suitable choice of the linear predictor, link function and distribution of dependent variable.

4.1.1 Link function

A link function forms a connection between the response of interest and the particular covariates observed. It was mentioned that the use of canonical link function leads to several convenient mathematical properties and some calculations then become easier. However, it does not mean that such link function will be usable, because it might not fit the data well or the interpretation of coefficients may be unreasonable or unexplainable. Due to these facts, commonly used link functions are log link, $g(\cdot) = \log(\cdot)$, or identity link function, $g(\cdot) = (\cdot)$. By using the log link, many statistical software packages require that the response in run-off triangle is positive. If a few negative values occur, log link can be applied, but the non-positive values must be replaced by the positive ones close to zero. The estimates vary widely due to slightly different choices of such values. Therefore, this approach of replacing values will not be applied in our case. In insurance practice, log link is preferred due to its interpretation, so we will use this link function as well. Nevertheless, only datasets with positive incremental data in run-off triangles will be shown.

4.1.2 Linear predictor

Due to the very specific structure of our data, there are not many choices for the linear predictor from (4). The simple one can be written as

$$\eta_{ii} = \beta_0 + b_i + \beta_t \,, \tag{7}$$

Where β_0 is an intercept, same for all accident years i. In order to avoid over-parameterization, β_1 is equal to zero. Random effect b_i can be explained together with β_0 as random intercept with mean equal to β_0 and represents the random effect of the accident year i. Finally, β_t captures the impact of change for particular development year. However, it is possible to use a more complex model to capture sophisticated covariance structure. In order to do this, random effects b_{it} for each development year t are included in the following equation

$$\eta_{ii} = \beta_0 + b_{i0} + \beta_t + b_{it} \,, \tag{8}$$

where $\beta_1 = b_{i1} = 0$ for all i and factor $\beta_t + b_{it}$ can be taken as random with mean β_t. Equation (4) can be written also as follows

$$\eta_{it} = \boldsymbol{X}_{it}^{T}\boldsymbol{\beta} + \boldsymbol{Z}_{it}^{T}\boldsymbol{b}_i \,,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_2, \ldots, \beta_p)^T$, random vector $\boldsymbol{b}_i = (b_{i0}, b_{i2}, \ldots, b_{in})^T$ and vector $\boldsymbol{X}_{it}$ is defined using dummy variables

$$X_{it} = (1, d_{2t}, \ldots, d_{nt})^{\mathrm{T}}.$$

Dummy variables are defined $d_{it} = 1$ for $i = t$ and zero otherwise. Vector $\boldsymbol{Z}_{it}$ equals to vector $\boldsymbol{X}_{it}$ as a consequence of given linear predictor.

The practical part of this paper deals only with the simplest model of linear predictor (7), because the one of the aims is to compare the suitability of the GLM and GLMM methods in a certain way.

4.1.3 Distribution of the incremental claims

We assume conditional distribution of incremental claims Y_{it} given $\boldsymbol{b}_i$ belongs to exponential family distribution. The choice of a suitable distribution is made according to the fitted values and residual diagnostics, where the relation between the mean and variance is investigated. The connection between mean and variance is described through variance function $V(\cdot)$ and for chosen widely-used distributions it is given by

$$Var(Y_{it}\,|\boldsymbol{b}_i) = \begin{cases} \varphi & \text{Gaussian distribution,where } V(\mu_{it}) = 1, \\ \varphi\mu_{it}^3 & \text{Inverse Gaussian distribution,where } V(\mu_{it}) = \mu^3{}_{it} \\ \varphi\mu_{it}^2 & \text{Gamma distribution,where V}(\mu_{it}) = \mu^2{}_{it} \end{cases}$$

In order to determine a suitable distribution, *Pearson residuals* are elaborated

$$r_{it}^{(P)} = \frac{Y_{it} - \hat{\mu}_{it}}{\sqrt{V\left(\hat{\mu}_{it}\right)}} \,.$$

A disadvantage of the Pearson residuals is that they are often markedly skewed. In order to "normalize" the residuals, *Anscombe residuals* were defined as

$$r_{it}^{(A)} = \frac{Y_{it} - \hat{\mu}_{it}}{\sqrt{V\left(\hat{\mu}_{it}\right)}}$$

where $A(\mu) = \int_{-\infty}^{\mu} V^{-\frac{1}{3}}(t)\mathrm{d}t$, which practically appear not to be skewed. Besides that, *deviance residuals*

$$r_{it}^{(D)} = sign(Y_{it} - \hat{\mu}_{it})\sqrt{d_{it}}$$

where d_{it} is the deviance for one observation (Bolker *et al*., 2009), can be used instead.

In cases, where we really could not decide about a suitable model according to the residual diagnostics or fitted values, information criteria can be used, cf. Bolker *et al*. (2009). The Akaike information criterion (AIC), the Bayesian information criterion (BIC), and other ones can serve this purpose. However, we try to avoid using such criteria because there is enough information in the residual diagnostic for this purpose.

4.2 GLM method for claims reserving

Application of GLM to claims reserving is also presented for the benchmark purposes. Just as for GLMM, a suitable linear predictor is discussed as well as link function. One of the differences between GLM and GLMM in claims reserving is that the effect of accident year is estimated in GLM unlike in GLMM where this effect can be predicted.

The specification of link function is the same as in the previous section. Log link is preferred due to its interpretation and practical usage in insurance. In order to compare the GLMM and GLM, same distributions of incremental claims are assumed, namely Gaussian, inverse Gaussian and gamma.

4.2.1 Linear predictor

Mean structure in the GLM is different than in the GLMM, due to the absence of random effects. As we mentioned, the choice of the linear predictor is a bit limited due to the interpretation and structure of claims data. Firstly, basic linear predictor, which use $2(n-1)+1$ unknown coefficients, is given by

$$\eta_{it} = \gamma + \alpha_i + \beta_t, \tag{9}$$

Where $\alpha_1 = \beta_1 = 0$ and α_i represents effect of accident year i, β_t effect of development year t. This can also be rewritten into vector notation

$$\eta_{it} = \boldsymbol{X}_{it}^{T}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\gamma, \alpha_2, \ldots, \alpha_n, \beta_2, \ldots, \beta_n)^{\boldsymbol{T}}$, and vector $\boldsymbol{X}_{it}$ is defined using dummy variables as $(1, d_{2i}, \ldots, d_{ni}, d_{2t}, \ldots, d_{nt})^{\boldsymbol{T}}$. There are several other linear predictors, *e.g.* Hoerl curve with the log link function, which can be parameterized by vectors $\boldsymbol{X}_{it}$ and $\boldsymbol{\beta}$ as

$$\boldsymbol{X}_{it} = \left(1, d_{2i}, \ldots, d_{ni}, 2 \times d_{2t}, \ldots, n \times d_{nt}, d_{2t} \times \log 2, \ldots, d_{nt} \times \log n\right)^T,$$

$$\boldsymbol{\beta} = (\gamma, \alpha_2, \ldots, \alpha_n, \beta_2, \ldots, \beta_n, \lambda_2, \ldots, \lambda_n)^T.$$

This leads to linear predictor

$$\eta_{it} = \gamma + \alpha_i + t\beta_t + log\ t.$$

However, we should realize that we have only n $(n + 1)/2$ observations and model with $3\ (n - 1) + 1$ parameters, which is not very useful for our purpose. Due to the higher number of parameters relative to the lower number of observations, as well as possibility to compare it to GLMM approach, model from (9) is used in the practical part, even though it still has a lot of parameters.

The impact of inflation on the development of claims was discussed in Section 3, but in cases where the market is stable and inflation does not affect the amount of claims much, simpler models should be taken into account, *e.g.* models with fewer numbers or even without accident year factors. Such models lead to better interpretation, the estimates become more precise and efficient, which are very important and practical properties.

5. Practical Application of Models

Prepared GLMM or GLM framework is applied on incremental claims Y_{it} from run-off triangle in Table 1, which represents known observations of random variables Y_{it}. The lower right part of the rectangle, Y_{it}, $i = 1, ..., n$, $n \geq t > n - i + 1$, is unknown and needs to be predicted in order to estimate the amount of total reserves which equals to

$$R^{(n)} = \sum_{i=2}^{n} R_i^{(n)}.$$

We have dealt with the GLM and GLMM approaches using unbalanced design, *i.e.* $t = 1, \ldots, n_i$ and $i = 1, \ldots, N$ where n_i stands for number of observations in i-th subject. Rewritten into reserving data structure (run-off triangles) $n_i = n - i + 1$ and $N = n$, where n is current year and i is accident year, which symbolizes i-th subjects. From now on, it holds that accident year (= subject) i is from the range $1, \ldots, n$.

As it was already pointed out, this practical part focuses on models with log link function and linear predictors in form (7) for GLMM. There are three proposed models for the GLMM: first one with Gaussian, second one with inverse Gaussian, and the last one with gamma distribution. These models are compared with three GLM models with the same distributions, same log link function, and linear predictor (9).

5.1 Database

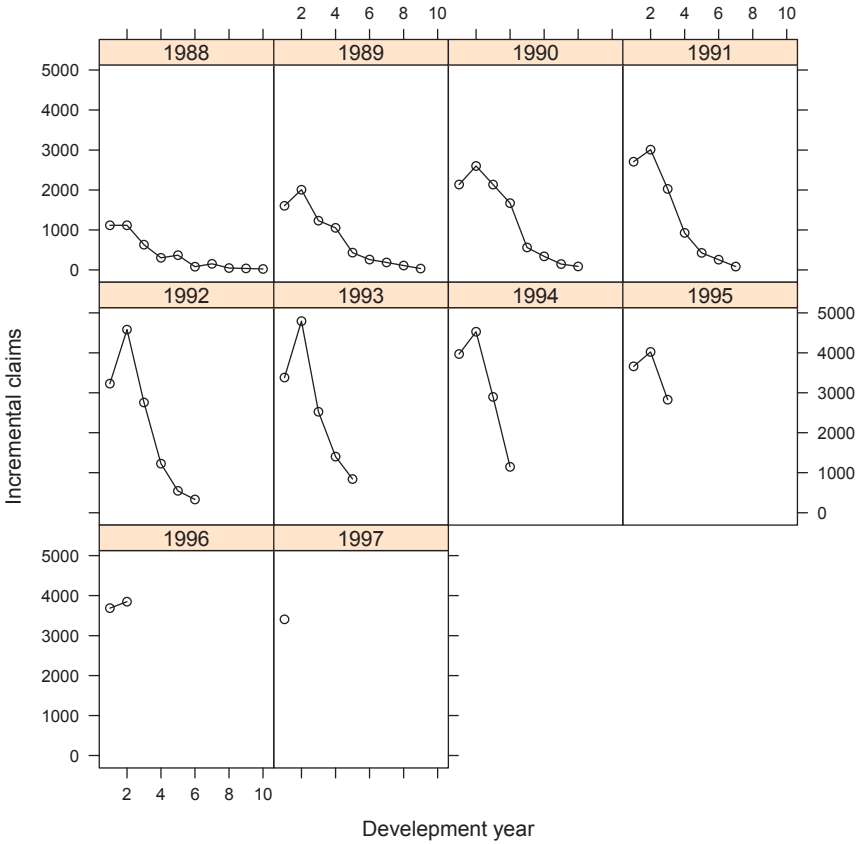
The results will be demonstrated on one dataset of claims from the National Association of Insurance Commissioners (NAIC) database, which can be downloaded from Meyers and Shi (2011). The database contains cleaned (*i.e.* deflated and large claims omitted) claims

developments of several lines for business for U.S. property casualty insurers. The data corresponds to claims from accident years 1988–1997 indexed from 1 to 10 with 10 years of development lag. Both upper and lower triangles are included, so we use upper triangle to develop the model and, consequently, to test its performance. Then, a retrospective analysis using all data including lower triangle is made as well.

5.2 GLMM *versus* GLM

Dataset by Hastings Mut, Ins. Co. from Workers' compensation line of business is chosen in order to demonstrate that the GLMM may be more suitable that the GLM. It can be seen from incremental claims in Figure 1, that the time lag between the claims occurrence and the claims payments in this dataset is considerably large. Possible explanation is that annuities are included in the data. The main amount of claims is paid in the second development year, not in the first one, as is usual.

Figure 1 | Claims Development for Each Accident Year



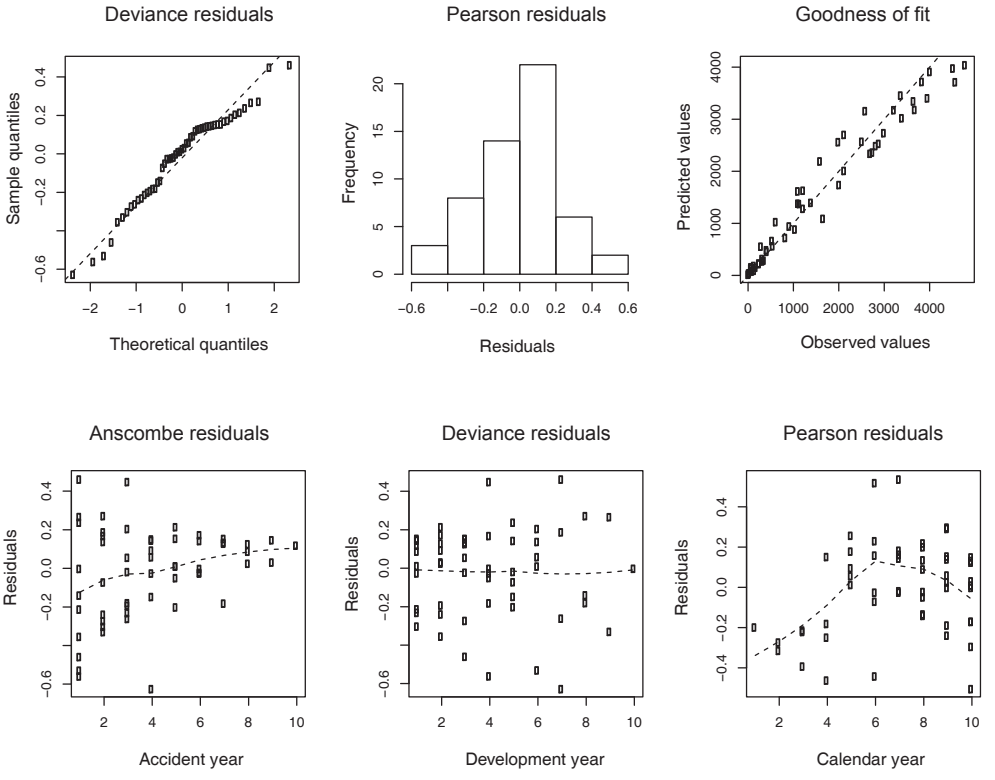
Source: Dataset by Hastings Mut. Ins. Co. from Workers' compensation line of business of the national Association of Insurance Commissioners (NAIC) database, which can be downloaded from Meyers and Shi (2011).

After a more detailed analysis of the data, it can be seen that the size of the peaks varies and no "trend" is present, *i.e.* a larger amount of claims in the first development year does not imply a proportionally bigger or lower peak. It is a very important property of the data and should be taken into account. This could be a case where the GLMM fits better, due to Bayesian approach used for prediction of the random accident year factors. This Bayesian property could handle such variation of the data properly. Similar behaviour appears in other analysed datasets, but this is the most representative one. It does not, however, mean that after such variation occurs, only GLMM should be chosen. Nevertheless, we should be aware of this during the model selection.

5.3 Residual diagnostic

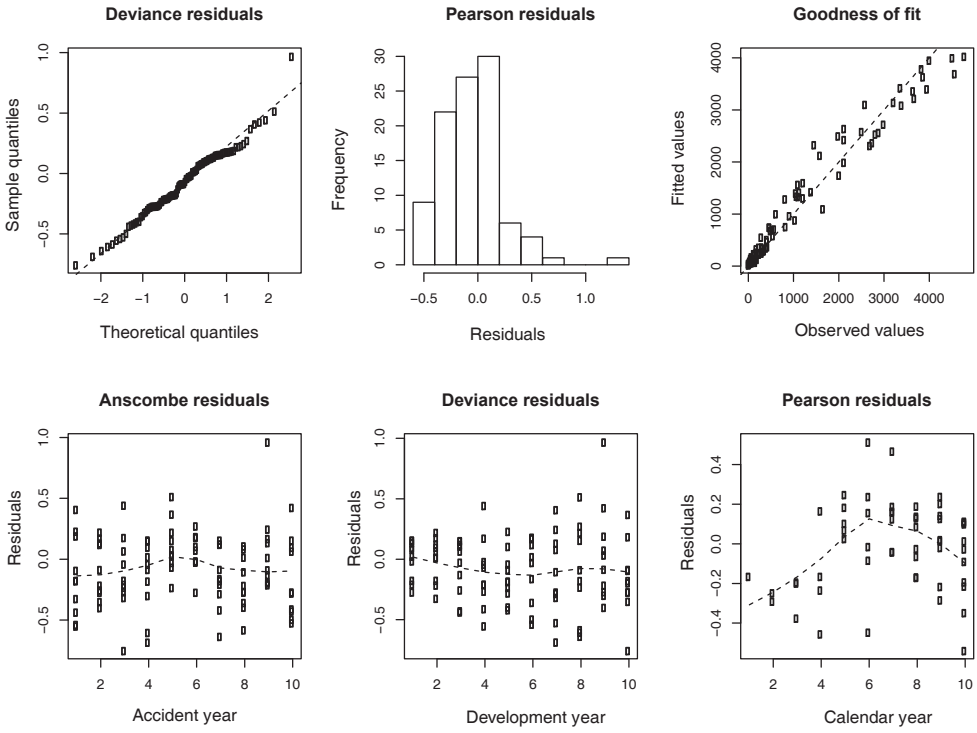
Firstly, all selected models are fitted on the upper triangle and then residual diagnostic based on the upper triangle is performed as well. Next, a comparison of all GLMM and GLM is made.

Figure 2 | Residual Diagnostics for the Gamma GLMM Generated Using the Upper Triangle



Source: see Figure 1.

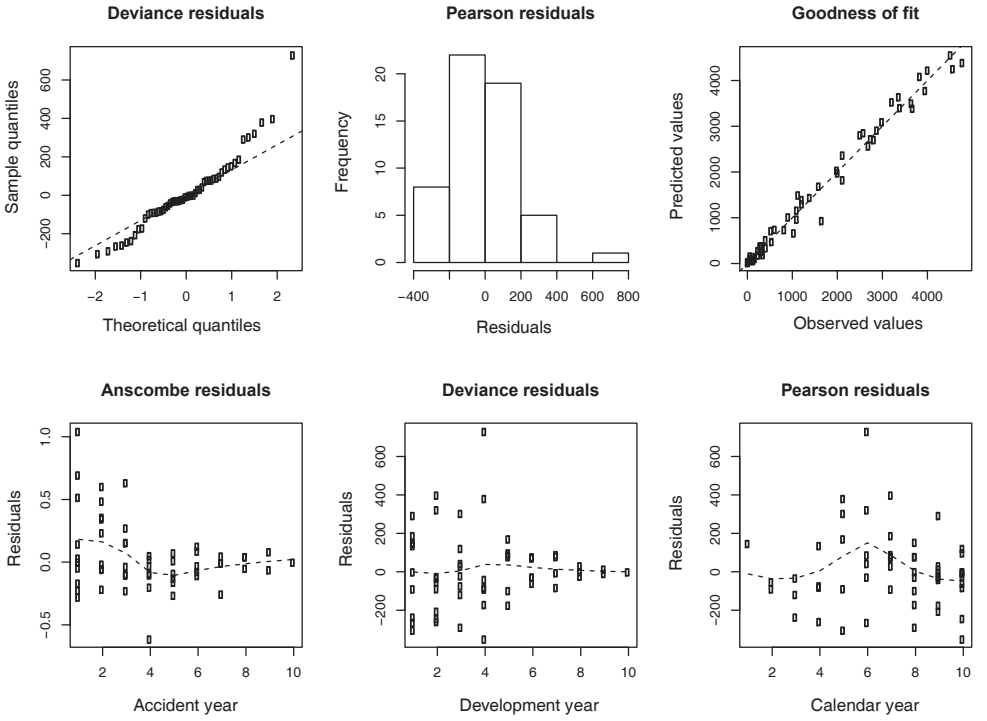
Figure 3 | Residual Diagnostics for the Gaussian GLMM Generated Using the Upper Triangle



Source: see Figure 1.

Based on the residual diagnostics in Figure 2 as well as suitability of the gamma distribution for the severity modelling (*i.e.* the support of the gamma distribution is only on the positive half of the real line) in comparison with other proposed GLMM, the gamma GLMM was chosen as the preferable one. The first diagnostic plot in Figure 2 is a QQ plot for the residuals, where points are expected to be placed near the diagonal line. In our case, we show the QQ plot for the deviance residuals, where we observe that the pairs of the theoretical and empirical residuals' quantiles are close to the dashed diagonal line. The situation for the Pearson as well as the Anscombe residuals is similar. The next plot displays a histogram of the Pearson residuals, which reveals that the empirical probability mass of the residuals is almost symmetrically placed around zero. Nevertheless, we suggest to pay more attention to the QQ plot compared to the histogram. The histogram is less preferable, because it may be visually misleading to interpret it in smaller sample sizes. We show the histogram just for the purpose that it is traditionally included in the residual diagnostics. The third subfigure is a plot of fitted values with respect to their observed counterparts, where all points should be and indeed are placed near the diagonal line. This implies that the fitted values mimic the original observed ones and, hence, the gamma model fits the data reasonably.

Figure 4 | Residual Diagnostics for the Gamma GLMM Generated Using the Whole Rectangle



Source: see Figure 1.

The remaining three subfigures show the relationship between the residuals (any type of residuals may be used) and the accident/development/calendar year. Here, the pay- ments for the same calendar year mean that these payments possess a constant sum of the corresponding accident and development year, *e.g.* they form diagonals in the run-off triangle of claim payments (Table 1). On the one hand, there is no visible pattern in case of the accident as well as the development year effect regardless of the type of residuals (dashed lines correspond to the loess curves), which implies that the gamma model is suitable. On the other hand, there is a slightly heterogeneous behaviour of the residuals with respect to the calendar year, but we could not reach a more reasonable fit from the remaining possible models.

Consequently, we perform residual diagnostics for the GLM with linear predictor from (9). Diagnostic figures were just as good as for the GLMM. However, we should take into account that we have to estimate $2n - 1$ unknown parameters in the GLM unlike $n + 1$ (n for development factors and one for the variance of the random effect b_i) in the GLMM. Based on our previous diagnostics, the mentioned behaviour of the data, and the number of parameters, which must be estimated, the *gamma GLMM* was chosen as the final one. For the sake of completeness, we provide the residual diagnostics for the Gaussian model in Figure 3, where it is visible that the residuals are more skewed and more heteroscedastic compared to the gamma model.

One can also inspect the residual diagnostics in Figure 4 based on the whole rectangle of incremental claims. All six subfigures look similarly as in Figure 2. The only small visual difference is the presence of one partially outlying observation coming from the lower triangle ("future" payments), which was not observed in the original upper run-off triangle to which our gamma model was fitted. Nevertheless, such a comparison of the overall goodness of fit taking into account the "future" data reveals that the gamma model is indeed a suitable one.

The next step after choosing an appropriate model is the prediction of reserves, which is described in the following section.

5.4 Predictions

Figure 5 illustrates all GLMM predictions for each accident year beginning directly after the vertical line. Fitted values are displayed before the vertical line. Based on the fitted values, the gamma model seems to be still the right choice. According to this figure, the Gaussian model looks suitable as well. Considering the predicted values, it is hard to say from Figure 5, which GLMM model has the most precise overall prediction. This fact is clearer from Table 2, where real occurred liabilities are computed using information from the lower triangle and compared with all the GLMM and GLM predicted reserves. Moreover, the chain ladder model is used for comparison as well. However, we should be aware of the fact that this model has different assumptions and, subsequently, different interpretation.

Table **2** shows that reserves predicted by the inverse Gaussian GLMM are the closest to the real observed liabilities despite the worst residual diagnostic and our chosen gamma GLMM was the second best in the prediction of reserves. All the GLM overestimate the occurred liabilities and almost all GLM predictions of the reserves are a lot higher than the GLMM predictions, except the Gaussian GLMM. Predictions provided by the GLMM are more precise for than the GLM predictions, even more precise than the chain ladder prediction in this case.

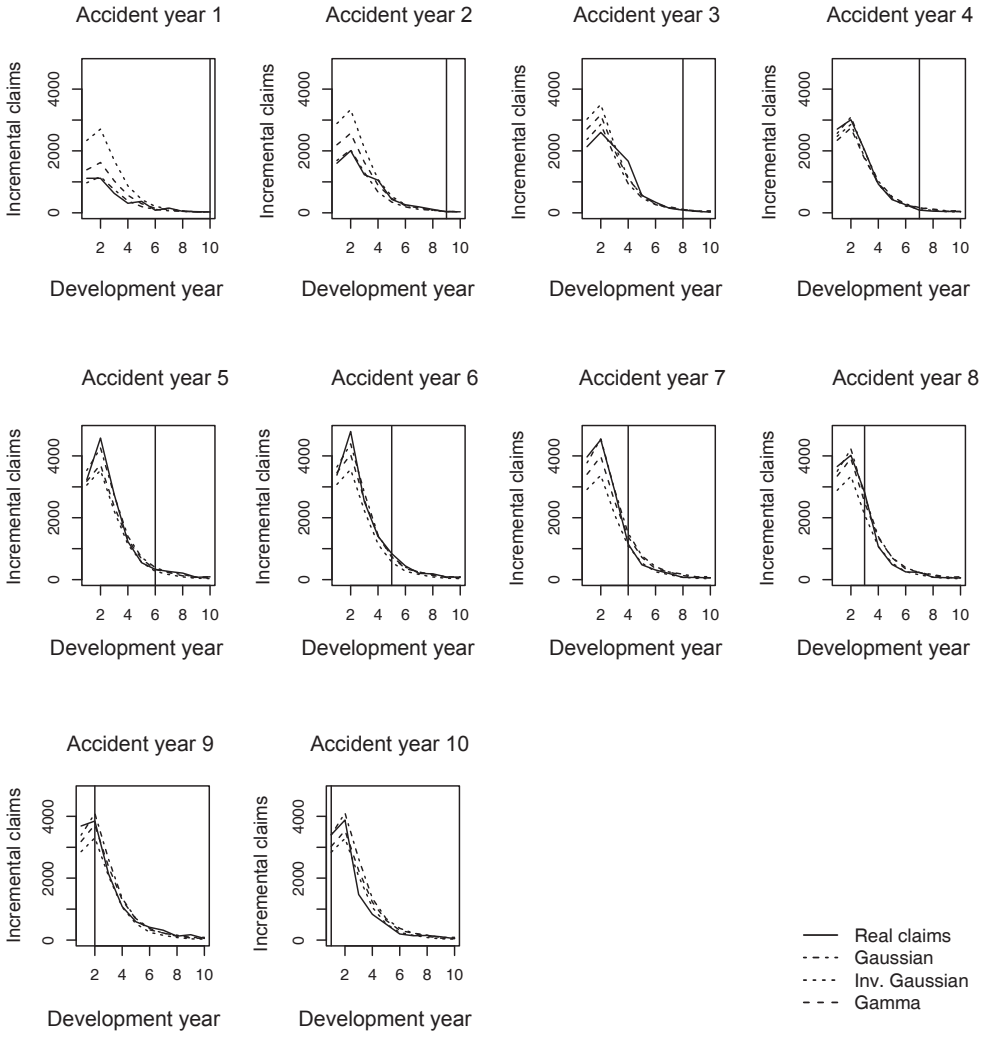
To sum up, according to the precision of the prediction, the residual diagnostics, the number of unknown parameters and the structure of claims development throughout accident years, the GLMM approach is more suitable for this dataset. Analogous situations occur for other data sets from the NAIC database, that the prediction of reserves is more accurate when using the GLMM instead of the traditional GLM.

Table 2 | Real and Predicted Reserves

Real	Mack	GLMM	Predictions	GLM	Predictions
17,475	22,625	Gaussian	22,033	Gaussian	22,033
		Inv. Gaussian	16,077	Inv. Gaussian	21,923
		Gamma	19,672	Gamma	22,659

Source: see Figure 1

Figure 5 | Predicted *versus* Real Claim Amounts



Source: see Figure 1.

6. Conclusions and Discussion

This paper proposes the *mixed effects* modelling technique as a suitable stochastic method for claims reserving. Classical stochastic methods usually assume only fixed (non-random) effects for the impact of the accident and development years. Therefore, these traditional approaches suffer from the lack of *flexibility* in predicting the future claim amounts compared to our approach allowing for the *random effects*. The GLMM applied for the run-off triangles yield *more precise prediction* of the reserves, because of *sparing the number of parameter*s for the fixed effects representing the accident years. The main contribution is

a novel technique for claims reserving together with the residual diagnostics for model suitability and selection.

A possible further extension of the proposed approach can be incorporating *more than one random effect* in the model for claims reserving. In that case, one can think about correlated random effects, which enables *modelling dependencies* between claim amounts. On the one hand, the prediction of reserves can become more precise by releasing the assumption of independent claim amounts. On the other hand, more information from the data needs to be spent for the estimation of additional random effects, which could lead to decrease of the prediction precision.

References

Antonio, K., Beirlant, J. (2007). Actuarial Statistics with Generalized Linear Mixed Models. *Insurance: Mathematics and Economics*, 40(1), 58–76, https://doi.org/10.1016/j.insmatheco.2006.02.013

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, H. H., White, J. S. (2009). Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution. *Trends in Ecology and Evolution*, 24(3), 127–135, https://doi.org/10.1016/j.tree.2008.10.008

Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S. (2002). *Analysis of Longitudinal Data.* 2nd Edition. New York: Oxford University Press.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models.* Boca Raton: CRC Press. England, P. D., Verrall, R. J. (2002). Stochastic Claims Reserving in General Insurance. *British Actuarial Journal*, 8(3), 443–518, https://doi.org/10.1017/s1357321700003809

Fitzmaurice, G. M., Laird, N. M., Ware, J. H. (2004). *Applied Longitudinal Analysis.* Boston: John Wiley and Sons.

Greene, W. H. (2002). *Econometric Analysis.* New Jersey: Prentice Hall.

Hudecová, Š., Pešta, M. (2013). Modelling Dependencies in Claims Reserving with GEE. *Insurance: Mathematics and Economics*, 53(3), 786–794, https://doi.org/10.1016/j.insmatheco.2013.09.018

Lehmann, E. L. (1983). *Theory of Point Estimation.* New York: Wiley.

McCulloch, C. E., Searle, S. (2001). *Generalized Linear and Mixed Models.* Boston: John Wiley and Sons.

Meyers, G. G., Shi, P. (2011). Loss Reserving Data Pulled from NAIC Schedule P. [Online; Updated September 01, 2011; Accessed June 10, 2014]. Available at: http://www.casact.org/research/index.cfm?fa=loss_reserves_data

Pešta, M., Okhrin, O. (2014). Conditional Least Squares and Copulae in Claims Reserving for a Single Line of Business. *Insurance: Mathematics and Economics*, 56(1), 28–37, https://doi.org/10.1016/j.insmatheco.2014.02.007

Rabe-Hesketh, S., Skrondal, A. (2002). Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature. *The Stata Journal*, 2(1), 1–21. Available at: http://ageconsearch.umn.edu/bitstream/115947/2/sjart_st0005.pdf

Rao, C. R., Toutenburg, H., Fieger, A., Heumann, C., Nittner, T., Scheid, S. (1999). *Linear Models: Least Squares and Alternatives.* 2nd Edition. Hoboken: Wiley Finance.

Raudenbush, S. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, 9(1), 141–157, https://doi.org/10.2307/1390617

Wüthrich, M. V, Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Hoboken: Wiley Finance.