Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS

Michal Gerthofer

Claims reserving within the panel data framework

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Michal Pešta, Ph.D.
Study programme: Mathematics
Specialization: Financial and Insurance Mathematics

Prague 2015

I would like to express my appreciation and sincere gratitude to my supervisor, RNDr. Michal Pešta, Ph.D. for his time, patience and all valuable materials which he provided me. His expert advices were very helpful to me.

I cannot find a word of thanks for my family, who was always there for me with infinite encouragement, love and care.

Last but not least, who I want to extend my thanks to, are my friends who encouraged and helped me whenever I needed.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague, 31st July 2015 Michal Gerthofer

Název práce: Rezervování škod v rámci panelových dat

Autor: Michal Gerthofer

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Předložená diplomová práce řeší problém závislosti mezi vysvětlovanými proměnnými v rámci jednotlivých subjektů v konceptu zobecněných lineárních modelů. Rezervování v neživotním pojištění má významný vliv na finanční postavení společnosti. Text práce představuje základní aktuárské pojmy, značení a metody. Hlavní část je zaměřena na modelování panelových dat, zejména na zobecněné lineární smíšené modely (GLMM), zobecněné odhadovací rovnice (GEE) a jejich použití v rezervování. Hlavním cílem práce je ukázat výhody, nevýhody, omezení a porovnání těchto přístupů na reprezentativních data setech, které byly vybrány na základě výsledků analýzy celé databáze. Významná pozornost je věnována procesu výběru modelu a prostředkům k tomu použitých. Získané výsledky jsou nakonec shrnuty v tabulkách, grafech a navzájem porovnány.

Klíčová slova: stochastické rezervování škod, panelová data, zobecněné smíšené lineární modely, zobecněné odhadovací rovnice

Title: Claims reserving within the panel data framework

Author: Michal Gerthofer

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In the presented thesis the issue of dependency between response variables within the subjects in the generalized linear models framework is investigated. Reserving in non-life insurance is a key factor for the financial position of a company. The text introduces the basic actuarial notation, terminology and methods. The main part is focused on panel data framework, especially Generalized Linear Mixed Models (GLMM) as well as Generalized Estimating Equations (GEE), and their application on claims reserving. The aim of this thesis is to show the advantages, disadvantages, limitations and the comparison of these approaches on representative datasets, which were chosen according to results obtained from whole database analysis. Significant focus is on model selection and diagnostics used for this purpose. Finally, the obtained results are summarized in tables, figures and the comparison of the methods is provided.

Keywords: stochastic claims reserving, panel data, generalized linear mixed models, generalized estimating equations

Contents

Introduction **3**

1 Panel data framework **5**

1.1 Brief historical overview 5
1.2 Characterization of the data structure 6
1.3 Basic Models 7
1.3.1 Pooled model (Classical linear model) 7
1.3.2 Error component models 7
1.3.3 Linear mixed models 9
1.4 Generalized Linear Model 11
1.5 Generalized Linear Mixed Model 15
1.5.1 Advantages of GLMM 15
1.5.2 Definition of GLMM 15
1.5.3 Estimation of parameters 16
1.6 Generalized Estimating Equations 17
1.6.1 Estimation of parameters 18
1.6.2 Correlation structure 19
1.6.3 Testing coefficients 21

2 Introduction to reserving theory **22**

2.1 Reserving terminology and notation 22
2.2 Basic reserving methods 23

3 Claims reserving in panel data framework **24**

3.1 GLMM method for claims reserving 24
3.1.1 Link function 24
3.1.2 Linear predictor 25
3.1.3 Distribution of incremental claims 25
3.2 GEE method for claims reserving 26
3.2.1 Linear predictor 26
3.2.2 Variance function 27
3.2.3 Correlation structure 27

4 Practical application of models **28**

4.1 Datasets 28
4.2 Claims reserving within GLMM 29
4.2.1 Dataset 29
4.2.2 Residual diagnostic 31
4.2.3 Claims reserving 33
4.3 Claims reserving within GEE 35
4.3.1 Dataset 35
4.3.2 Residual diagnostic 37
4.3.3 Claims reserving 39
4.3.4 Testing coefficient 41
4.4 Problematic example 44

4.5 Comparison of results 48

Conclusion **49**

Bibliography **50**

List of Figures **50**

List of Abbreviations **53**

A Source code **54**

A.1 Main control script 54

A.2 GLMM script 57

A.3 GEE script 59

Introduction

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> *George E.P. Box*

The claims reserving problem or the run-off problem, has been one of the most important issues handled in general insurance for many years. The proper determination of the reserves amount has a key impact on the financial position of an insurance company, especially in the non-life insurance. Many various methods have been developed for this purpose beginning with deterministic approaches such as an original chain-ladder and later stochastic models have been proposed.

However, various approaches and models lead to different results. Every method is based on different ideas and uses different principles. Therefore, the most important question in the practical application of these methods is model selection and the adequacy of chosen models.

Stochastic methods for modelling claims development became popular among practitioners in order to use more information about the data and consequently, to get deeper detail about the estimate as well. The major part of these approaches requires independency of the incremental claims. In practice, this assumption often does not hold and the methods can provide misleading results. This thesis deals with the stochastic models based on generalized linear models, especially Generalized Linear Mixed Models (GLMM) and Generalized Estimating Equations (GEE), which are able to handle the aforementioned dependency.

The purpose of the thesis will be the construction of suitable models for claims reserving framework, then a description of model selection on representative datasets and subsequently, the advantages and disadvantages of adequate models will be discussed. During this process, special attention is paid to the detailed analysis of a huge number of datasets and their behaviour.

The structure of the thesis is as follows. In the first chapter, panel data framework is introduced in general, starting with the most basic linear models and continuously, the correlation is introduced into models with random effect. Then, generalized linear models are presented with their strength in various mean structure and possible distributions of random variables. Finally, the advantages of these two mentioned models are introduced in GLMM and GEE models. For the purpose of the simplification of GEE models, testing of coefficient is discussed as well.

The aim of the second chapter is to introduce the claims reserving problem in non-life insurance and its standard notation. The incremental or cumulative claims are understood as random variables ordered in the development triangles. At the end of the chapter, basic reserving methods are introduced.

Next, the focus lies on the application of the panel data theory in claims reserving framework. Specification of the models, which satisfy the theory associated with the actuarial practice, is discussed. In order to choose the most adequate model, useful residual properties are presented. Moreover, the reasons for simpler models, if it is possible, are listed.

Finally, in the fourth chapter, the application of model selection and residual diagnostic on real-life data is executed using `R` software. Firstly, the analysis on the whole database is made and subsequently three main representative datasets are chosen in order to demonstrate their strengths and weaknesses, as well as the differences in the considered approaches.

Notation agreement

In this text we use following notation for indices, $\mathbf{Y}_{it} \equiv \mathbf{Y}_{i,\,t}$ and $\mathbf{Y}_{itj} \equiv \mathbf{Y}_{i,\,t,\,j}$. 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 $\mathbf{Y}_{1+it+n} \equiv \mathbf{Y}_{1+i,\,t+n}$. The same holds for numbers in indices due to the fact that we only use integers less than 10 in indices, thus notation is given by $\mathbf{Y}_{12} \equiv \mathbf{Y}_{1,\,2}$ or $\mathbf{Y}_{1n-1} \equiv \mathbf{Y}_{1,\,n-1}$.

1. Panel data framework

1.1 Brief historical overview

The first mention of panel data framework is dated in the second half of the 19th century. British astronomer George Biddel Airy laid the foundations for the linear mixed-model formulation (1861), which he applied to errors of observation in astronomy.

About fifty years later, it was put on a more formal theoretical footing in the seminal work of R. A. Fisher (1918), where he defined the terms "variance" and "Analysis of variance" (ANOVA). In his later works, he elaborated on the concept of models with fixed and random effects.

It did not take a long time until statisticians recognized similarities between a (panel data) structure with N individuals and T repeated measurements and data collected in randomized blocks. Thus, it seemed natural to apply ANOVA methods developed later (e.g., Yates, 1935; Scheffe, 1959) to the repeated-measures data collected from studies of panel data, where the individuals were considered as the blocks.

Later, models with random effects were suggested for no experimental data, e.g., in astronomy, econometric or biostatistics. On the other hand, models with fixed effects should be used for no experimental data, where there are concrete procedures like in industry or agriculture.

The analysis of change is a fundamental component of many research endeavours. Thus, these methods have been gradually used in almost every discipline including econometrics, biostatistics, pharmacy, insurance, industry, agricultural etc.

For example, one of the first authors of econometrics application was Irving Hoch (1962), who was estimating the Cobb-Douglas production function for 6 years of data for 63 farms in Minnesota. Another improvement in this field was in the research of dynamic models where models contain lagged explanatory variables or dependent variables. This fact is very reasonable, especially for econometric data because there is usually a very strong correlation to previous observations.

Another huge application was made in the analysis of human investigation, which expanded explosively in the second half of the twentieth century by the US government through the legislative foundation for the modern National Institute of Health (NIH). In this field of research, term longitudinal or clustered data is often used instead of panel data and term individual or panel member is usually replaced by subject or cluster. In this text, we are going to deal with longitudinal data (time ordered data within subjects) which is a special case of more general defined clustered data.

Investigators in this field were interested in the treatment of diseases that are not typically life threatening and wanted to understand the development of disease and to identify factors that cause changes. Before they began investigating temporal patterns of change, new and more computationally sophisticated approaches had to be invented. In order to do this in the early 1980s, Laird and Ware proposed the use of the Expectation Maximization (EM) algorithm to fit a class of Linear Mixed Model (LMM) and Jennrich and Schluchter (1986) proposed a variety of alternative algorithms, including Fisher scoring (IWLS) and Newton-Raphson algorithms. Ten years later, Liang and Zeger introduced Generalized Estimating Equations (GEE) and proposed a family of Generalized Linear Model (GLM) for fitting repeated observation of binary and counted data.

For more details about the history, see Fitzmaurice et al. [2004, Chapter 1.1] or Diggle et al. [2002, Chapter 1].

1.2 Characterization of the data structure

Panel data contains observations of multiple phenomena obtained over multiple time periods for the same individuals or subjects. Time series and cross-sectional data are special cases of panel data that are in one dimension only. 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.

Before we start describing models in mathematical formulae, we present some advantages of panel data. One of them is the large number of observations, which can bring a larger number of degrees of freedom and reduce the collinearity among explanatory variables. More importantly, longitudinal data allows a researcher to analyse a number of important questions that cannot be answered using cross-sectional or time-series data sets, e.g., by using models with fixed effects we can study how big a part of disturbances should be explained by individual effects. Disadvantages include challenging data collection and often an insufficient series length.

In this part, panel data framework is described more mathematically. Firstly, the commonly used basic linear model is defined by the following form

$$Y_{it} = X_{1it}\beta_1 + X_{2it}\beta_2 + \cdots + X_{pit}\beta_p + u_{it}, \tag{1.1}$$

where $p \in \mathbb{N}$, $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots, Y_{1T}, \ldots, Y_{NT})^\top$ is the outcome of interest, vector of explanatory variables, $\mathbf{X}_{it} = (X_{1it}, \ldots, X_{pit})^\top$, is part of the model matrix

$$\mathbf{X} = (\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1T}, \ldots, \mathbf{X}_{NT})^\top,$$

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression parameters, $\mathbf{U} = (u_{11}, \ldots, u_{1T}, \ldots, u_{NT})^\top$ is the vector of residuals (disturbances). In this part $t = \{1, \ldots, T\}$ is considered as time index and $i = \{1, \ldots, N\}$ is considered for individual or panel subject and this structure is called balanced design, i.e., the same number of observations for each subject.

The equation from 1.1 can be rewritten in the following form

$$Y_{it} = \mathbf{X}_{it}^\top \boldsymbol{\beta} + u_{it}, \tag{1.2}$$

where matrix notation is equal to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}.$$

In this text, we consider $\mathbf{X}$ as fixed but there is another approach, where $\mathbf{X}$ is taken as random. However, both of them lead to the same conclusions.

1.3 Basic Models

This section describes basic approaches for modelling panel data. Firstly, models without correlation structure between dependent variables are presented and secondly, correlation is introduced in basic form which is the main purpose of this section.

1.3.1 Pooled model (Classical linear model)

The first model presented is the simplest, the so called pooled model. It assumes that residuals u_{it} from (1.1) are not structured, which means that no individual or time effects are present within them and they are not correlated with explanatory variables. Moreover they are independent, identically distributed random variables with zero mean and finite, positive variance equal to σ_u^2. Therefore, there is no need to have panel structured data and components of $\mathbf{Y}$. Thus, $\mathbf{Y}$ and $\mathbf{U}$ can be equally reordered.

Finally, we can estimate unknown parameters applying the Ordinary Least Squares (OLS) method, which leads to the solution

$$\widehat{\boldsymbol{\beta}}_P = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y},$$

when an inverse matrix exists. For more information see, e.g., Rao et al. [1999, Chapter 3]. It is common that $\widehat{\boldsymbol{\beta}}$ depends on the number of observations, which is usually written as $\widehat{\boldsymbol{\beta}}(n)$, where n stands for the number of observations.

1.3.2 Error component models

Now we continue to more sophisticated models with structured residuals, so called error component models. These approaches allow us to model unobserved heterogeneity, e.g., we can add different intercepts for all individuals and thus better model dependent variables.

Error component models are split into one-way error component models and two-way error component models. The first mentioned models assume that residuals consist of classical residuals ,i.e., independent random variables with zero mean and constant positive variance, and cross sectional or time effect. The second mentioned models assume that residuals consist of all three elements (classical residuals, cross sectional and time effect).

One-way error component models

This text focuses only on one-way error component models where residuals consist of classical residuals and individual effects. Firstly, a model with fixed individual effects is presented, which as we mentioned before should be used for data with a smaller number of observed individuals. Next, attention is paid to a model with random effects which should be used for data with a huge number of individuals or subjects.

Form from (1.2) holds, but the breakup of residuals is added

$$Y_{it} = \mathbf{X}_{it}^\top\boldsymbol{\beta} + u_{it}, \qquad u_{it} = b_i + \varepsilon_{it}, \tag{1.3}$$

where individual random or fixed effects b_i reflect unobserved or unobservable factors that make individuals respond differently. Elements ε_{it} stand for the rest of unexplained residual effects that vary with time and individuals, i.e., classical residuals.

One-way error component models with fixed effect

Let's begin with the fixed effect model, where equation (1.3) can be re-written applying dummy variables as follows

$$Y_{it} = \mathbf{X}_{it}^{\top}\boldsymbol{\beta} + d_{1it}b_1 + d_{2it}b_2 + \cdots + d_{Nit}b_N + \varepsilon_{it},$$

where $t = \{1,\ldots,T\}$, $i = \{1,\ldots,N\}$ and d_{kit} is

$$d_{kit} = \begin{cases} 1 & i = k, \\ 0 & \text{otherwise.} \end{cases}$$

It can also be written in matrix notion as follows

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}_b\mathbf{b} + \varepsilon,$$

where $\mathbf{D}_b = \mathbf{I}_N \otimes \mathbf{1}_T$ is $NT \times N$ model matrix , see definition in Abbreviations (p. 53) and $\varepsilon = (\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{1T}, \ldots \varepsilon_{NT})^{\top}$ is the vector of classical residuals. Here, OLS method, in this case also called the Least Squares Dummy Variables (LSDV) method, can be applied to get the estimation of unknown vector of regression parameters
$\boldsymbol{\beta}_b = (\beta_1, \ldots, \beta_p, b_1, \ldots, b_N)^{\top}$ in the following form

$$\widehat{\boldsymbol{\beta}}_{dummy} = \left(\mathbf{X}|\mathbf{D}_b^{\top}\mathbf{X}|\mathbf{D}_b\right)^{-1}\mathbf{X}|\mathbf{D}_b^{\top}\mathbf{Y},$$

which can be expressed when an inverse matrix exists. For the definition of matrix $\mathbf{X}|\mathbf{D}_b$ see Abbreviations (p. 53).

It is worth mentioning that by using dummy variables for creating model matrix $\mathbf{X}|\mathbf{D}_.$, it is possible to model two-way error component models with both fixed, cross sectional and time effect as well.

One-way error component models with random effect

Until now, no correlation structure between dependent variables was considered. In a model with random effects, simple correlation structure is introduced. Equation (1.3) still holds, however with the following assumptions

$$\begin{aligned} &Y_{it} = \mathbf{X}_{it}^{\top}\boldsymbol{\beta} + b_i + \varepsilon_{it}, \\ &b_i \sim iid\left(0, \sigma_b^2 > 0\right), \\ &\varepsilon_{it} \sim iid\left(0, \sigma_\varepsilon^2 > 0\right), \\ &\mathsf{E}\left(\varepsilon_{it}b_j\right) = 0, \quad \forall\, i, j \text{ and } t, \\ &\mathsf{E}\left(b_ib_j\right) = 0, \quad i \neq j, \end{aligned} \tag{1.4}$$

where $r_i \sim iid\,(m,v)$ means that r_i are independent, identically distributed random variables with mean m and variance equal to v, for all i.

Presence of two elements within residuals implies the following relation

$$\begin{aligned} Cov\left(u_{it}, u_{is}\right) &= Cov\left(\varepsilon_{it}+b_i, \varepsilon_{is}+b_i\right) \\ &= Var\left(b_i\right)+Cov\left(\varepsilon_{it}, \varepsilon_{is}\right) = \\ &= \begin{cases} \sigma_b^2 & t \neq s,\ \forall i, \\ \sigma_b^2+\sigma_\varepsilon^2 & t = s,\ \forall i. \end{cases} \end{aligned}$$

Hence correlation is equal to

$$Corr\left(u_{it}, u_{js}\right) = \begin{cases} \frac{\sigma_b^2}{\sigma_b^2+\sigma_\varepsilon^2} & t \neq s,\ i = j, \\ 1 & t = s,\ i = j, \\ 0 & \text{otherwise.} \end{cases}$$

This model was presented in order to realize what causes correlation. This idea is very crucial for understand the following sections.

1.3.3 Linear mixed models

Finally, we come to generalization of one-way error component models with random effect, Linear Mixed Models (LMM), which allow us to model more complex covariance structures and can also handle the complications of mistimed and incomplete measurements in a very natural way. Model is given by

$$\begin{aligned} &Y_{it} = \mathbf{X}_{it}^\top \beta + \mathbf{Z}_{it}^\top \mathbf{b}_i + \varepsilon_{it}, \quad \text{or} \\ &\mathbf{Y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i + \varepsilon_i \end{aligned} \tag{1.5}$$

where $t = \{1, \ldots, n_i\}$, $i = \{1, \ldots, N\}$, i.e., different number of observations through subjects, the so called unbalanced design. Furthermore, $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots Y_{in_i}\right)^\top$ is vector of response for i-th individual and

$$\mathbf{Z}_i = \left(\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, \ldots, \mathbf{Z}_{in_i}\right)^\top$$

is a particular subset of model matrix

$$\mathbf{X}_i = \left(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{in_i}\right)^\top.$$

Next, random vector $\varepsilon_i = \left(\varepsilon_{i1}, \ldots \varepsilon_{in_i}\right)^\top$ is independent with vector $\mathbf{b}_i = \left(b_1, \ldots, b_q\right)^\top$ and both have multinomial normal distribution with zero mean and variance matrices $\mathbf{R}_i$, $\mathbf{G}$ respectively. Additionally, we assume $\mathbf{R}_i = \sigma_\varepsilon^2 \mathbf{I}_{n_i}$ because it is not possible to estimate unstructured both $\mathbf{G}$ and $\mathbf{R}_i$.

As we mentioned, one-way error component model with random effect, (1.4), is a special case of LMM design, (1.5), where $\mathbf{Z}_{it}^\top$ equals one and $\mathbf{b}_i$ is one dimensional random variable and changes a fixed intercept to a random one.

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

$$\begin{aligned} Cov\left(\mathbf{Y}_i\right) &= \Sigma_i \\ &= Cov\left(\mathbf{Z}_i \mathbf{b}_i\right)+Cov\left(\varepsilon_i\right) \\ &= \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i^\top + \mathbf{R}_i. \end{aligned}$$

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

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{Y}_i \right). \tag{1.6}$$

This is Best Linear Unbiased Estimate (BLUE) of $\boldsymbol{\beta}$. For more details and properties of the estimate, see Rao et al. [1999, Chapter 4] or Fitzmaurice et al. [2004, Chapter 4.2].

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

$$\widehat{\boldsymbol{\beta}}_{FGLS} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \widehat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \widehat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{Y}_i \right).$$

See more about FGLS in Greene [2002, Chapter 10.5].

Iterative Generalised 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 $\widehat{\boldsymbol{\Sigma}}_i$. This process is repeated until the required precision is obtained.

In many applications, inference is focused 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 $\mathbf{b}_i$, given the vector of response $\mathbf{Y}_i$, $(\widehat{\boldsymbol{\beta}})$. Using properties of join multivariate normal distribution, it can be written as

$$\mathsf{E}\left(\mathbf{b}_i | \mathbf{Y}_i\right) = \mathbf{G}\mathbf{Z}_i^\top \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{Y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right).$$

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

$$\widehat{\mathbf{b}}_i = \widehat{\mathbf{G}}\mathbf{Z}_i^\top \widehat{\boldsymbol{\Sigma}}_i^{-1} \left(\mathbf{Y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right).$$

Predicted response profile is then given by

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

which can be re-written as follows

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

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 $\mathbf{R}_i$ is large relatively to the between-subject variability

Σ_i, more weight is given to $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ than to the i-th observed response. For more details about these estimates see Fitzmaurice et al. [2004, Chapture 8.7].

The reason why it is better to use REML for estimation of Σ_i is that the ML method treats $\boldsymbol{\beta}$ as fixed but unknown quantities when the variance components are estimated. However, it does not take into account the degrees of freedom lost by estimating these fixed effects. This causes the ML estimator to be more biased than the REML estimate of Σ_i. It should be noted that the difference between the ML and REML estimation becomes less important when the number of sample size, N, is substantially larger than the dimension of $\boldsymbol{\beta}$. The advantage of ML over REML estimate of Σ_i is that it is possible to compare two models in terms of their fixed and random effects. On the other hand, if REML is used to estimate the parameters, it is possible to compare only models that are nested in their random effects terms and the same in their fixed effects. For more information about REML, see in Fitzmaurice et al. [2004, Chapter 4.4, 4.5].

1.4 Generalized Linear Model

Until now, all mentioned models work with dependent variables whose mean is defined on $\mathbb{R}$ and in LMM the dependent variable is expected to be normally distributed. This assumption is very restrictive for real data. Thus, we go on to more complex models which are able to handle dependent variables from various distributions and ranges of mean. Generalized Linear Models (GLM) deal with responses whose distribution functions belong to exponential family. This fact allows us to model, e.g., zero-one (alternative) dependent random variable. Before we focus on GLM in detail, it is appropriate to define exponential family of distributions.

Exponential family of distributions

Exponential family contains distributions with densities that can be written as

$$f(Y|\theta,\varphi) = \exp\left\{\frac{Y\theta - b(\theta)}{\varphi} + c(Y,\varphi)\right\}, \tag{1.7}$$

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

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

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

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

$$\begin{aligned} \mathsf{E}(Y) &= b'(\theta)\ (<\infty), \\ Var(Y) &= \varphi b''(\theta)\ (<\infty). \end{aligned} \tag{1.8}$$

Corollaries from (1.7) ($\varphi > 0$) and (1.8) ($Var(Y) > 0$) are that $b(\cdot)$ is convex function and $b'(\cdot)$ is strictly increasing. Hence $b'(\cdot)$ has a well-defined inverse.

Here we define variance function $V(\cdot)$ for which $Var(Y) = \varphi V(\mu)$ holds and satisfies $b''(\theta) = V[b'(\theta)]$. For more details, see Fitzmaurice et al. [2004, Chapter 10.5].

Estimation of parameters of Exponential family distribution

Let $Y_1, \ldots, Y_n$ be a random sample from the distribution with density (1.7), then we can use ML method for estimation of θ. Thus, we obtain score statistic

$$U_n(\theta|Y_1, \ldots, Y_n) = \frac{1}{\varphi} \sum_{i=1}^{n} \left[Y_i - b'(\theta)\right].$$

Then estimate is defined as solution of equation $U_n(\widehat{\theta}|Y_1, \ldots, Y_n) = 0$, which is equal $\widehat{\theta} = (b')^{-1}(\sum_{i=1}^{n} Y_i/n)$, where $\widehat{\theta} \equiv \widehat{\theta}(n)$.

Solution is unique because $b(\cdot)$ is convex and it does not depend on φ. Moreover, if regularity conditions are satisfied, the estimate is consistent and asymptotically normal

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}\left(0, \mathbf{I}^{-1}(\theta)\right), \tag{1.9}$$

where $\mathbf{I}^{-1}(\theta)$ is information number equal $\varphi/b''(\theta)$.

In cases where the dispersion parameter is unknown, the asymptotic variance of $\widehat{\theta}$ may change. However, the join information matrix for the vector $(\theta, \varphi)^\top$ from 1.10 is diagonal.

$$\mathbf{I}(\theta, \varphi) = -\mathsf{E}\, \frac{\partial^2 \log f(Y; \theta, \varphi)}{\partial(\theta, \varphi)\partial(\theta, \varphi)^\top} = \begin{pmatrix} I_{\theta\theta} & I_{\theta\varphi} \\ I_{\theta\varphi} & I_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} b''(\theta)/\varphi & 0 \\ 0 & I_{\varphi\varphi} \end{pmatrix} \tag{1.10}$$

Thus, corollary from (1.9) and (1.10) is that the asymptotic variance of $\widehat{\theta}$ given by $\varphi/b''(\theta)$ holds also when φ is unknown, which imply asymptotic independence of ML estimates $\widehat{\theta}$ and $\widehat{\varphi}$, when regularity conditions for vector $(\theta,\varphi)^\top$ are satisfied. Unknown parameter φ is not needed for estimation of θ, only if we are interested in asymptotic variance of θ. Then it can be estimated using moment estimator because ML estimate cannot often be calculated explicitly. For more detailed theory, see Lehmann [1983, Chapter 6.4]

Definition of Generalized Linear Model

Here we go on to the definition of GLM, where we want to express the dependence of $\mu_i \equiv \mathsf{E}\, Y_i$ on explanatory variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ using a more general model than the linear one.

GLM is given by the following conditions.

- $Y_1, \ldots, Y_n$ are independent random variables and distribution of Y_i depends on $\mathbf{X}_i$ through regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

- The canonical density of Y_i is the same as (1.7), except for the canonical parameter θ, which depends on $\mathbf{X}_i$ and $\boldsymbol{\beta}$. Therefore, from now on, we add index to this, i.e., $\theta_i \equiv \theta$. Additionally, $b(\cdot)$ is assumed to be twice differentiable for the reasons mentioned above.

- Dependence between $\theta_i, \mathbf{X}_i$ and $\boldsymbol{\beta}$ is expressed through *linear predictor*

$$\eta_i \equiv \mathbf{X}_i^\top \boldsymbol{\beta}.$$

- There is a known strict monotone, twice continuously differentiable function $g(\cdot)$ called *link function*, for which $g(\mu_i) = g\left[\mathsf{E}(Y_i)\right] = g\left[b(\theta_i)\right] = \eta_i$ holds.

It is useful to define *canonical link* as a link function which satisfies the following equation

$$g(\mu_i) = \eta_i = \theta_i = \mathbf{X}_i^\top \boldsymbol{\beta},$$

which implies $(b')^{-1}(\cdot) = g(\cdot)$. Then, applying first derivation on $g\left[b'(\theta_i)\right] = \theta_i$ we obtain $g'\left[b'(\theta_i)\right] b''(\theta_i) = 1$ and continuously $g'(\mu_i) = 1/V(\mu_i)$.

According to previous conclusions, the following parametrization is used

$$\begin{aligned}
\eta_i &= \mathbf{X}_i^\top \boldsymbol{\beta}, \\
\eta_i = g(\mu_i) &\Longleftrightarrow \mu_i = g^{-1}(\eta_i), \\
\mu_i = b'(\theta_i) &\Longleftrightarrow \theta_i = (b')^{-1}(\mu_i), \\
\eta_i = g\left[b'(\theta_i)\right] &\Longleftrightarrow \theta_i = (b')^{-1}\left[g^{-1}(\eta_i)\right].
\end{aligned}$$

Maximum Likelihood Estimation in GLM

Let $Y_1, \ldots, Y_n$ satisfy conditions of generalized linear model and φ is known dispersion parameter. Unknown parameter $\boldsymbol{\beta}$ can be estimated using ML method. Firstly, log-likelihood is given by

$$\ell_n(\boldsymbol{\beta}|Y_1,\ldots,Y_n) = \sum_{i=1}^{n} \left[\frac{Y_i\theta_i - b(\theta_i)}{\varphi} + c(Y_i,\varphi)\right],$$

where $g(\mu_i) = \mathbf{X}_i^\top \boldsymbol{\beta}$ and $\mu_i = b'(\theta_i)$. Furthermore, we express score function

$$\begin{aligned}
U(\boldsymbol{\beta}|Y_i) &= \frac{\partial}{\partial \boldsymbol{\beta}} \frac{Y_i\theta_i - b(\theta_i)}{\varphi} \\
&\overset{\text{chain rule}}{=} \frac{\partial}{\partial \theta_i} \frac{Y_i\theta_i - b(\theta_i)}{\varphi} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} \\
&= \frac{Y_i - \mu_i}{\varphi} \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} \mathbf{X}_i.
\end{aligned}$$

This leads to score statistic

$$U_n(\boldsymbol{\beta}|Y_1,\ldots,Y_n) = \frac{1}{\varphi} \sum_{i=1}^{n} w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i,$$

where $w(\mu_i) = 1/\left\{V(\mu_i)\left[g'(\mu_i)\right]^2\right\}$ (> 0) is weight function. As a consequence of this, the set of p likelihood equations for $\boldsymbol{\beta}$ can be expressed as follows

$$\sum_{i=1}^{n} w(\widehat{\mu}_i) g'(\widehat{\mu}_i)(Y_i - \widehat{\mu}_i)\mathbf{X}_i = \mathbf{0}, \tag{1.11}$$

where $\widehat{\mu}_i = g^{-1}(\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}})$.

If $g(\cdot)$ is canonical link function, then $g'(\cdot) = 1/V(\cdot)$, $w(\mu_i) = V(\mu_i)$, and consequently $w(\mu_i)g'(\mu_i) = 1$. Thus, score static and likelihood equations can be written in the following form

$$\begin{aligned} U_n(\boldsymbol{\beta}|Y_1,\dots,Y_n) &= \frac{1}{\varphi}\sum_{i=1}^{n}(Y_i - \mu_i)\mathbf{X}_i, \\ \sum_{i=1}^{n} Y_i\mathbf{X}_i &= \sum_{i=1}^{n}\widehat{\mu}_i\mathbf{X}_i, \end{aligned}$$

where $\widehat{\mu}_i = g^{-1}\left(\mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}\right)$ are called the fitted values.

If we want to obtain a unique solution of likelihood equations, then link function $g(\cdot)$ must be canonical and model matrix $\mathbf{X} = (\mathbf{X}_1,\dots,\mathbf{X}_n)^\top$ must have full column rank p, because otherwise, the equations may have multiple solutions. These conditions imply that log-likelihood is a concave function of $\boldsymbol{\beta}$, which can be proved by properties of Fisher information.

Here, we must use a numerical algorithm to solve the likelihood equations, due to the non-linearity in equations. They may iterate slowly and what is worse, they may converge at the wrong solution if there is no canonical link or model matrix does not have full column rank p.

Iterative Weighted Least Squares

We present one numerical method called Iterative Weighted Least Squares (IWLS), which is a special use of the Fisher scoring algorithm. Before we describe the algorithm it is necessary to state that the ML estimate, $\widehat{\boldsymbol{\beta}}$, in GLM solves the system of equations

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{Z}}\right), \tag{1.12}$$

where

$$\begin{aligned} \widehat{\mathbf{W}} &= diag\left[w(\widehat{\mu}_1),\dots,w(\widehat{\mu}_n)\right], \\ \widehat{\mathbf{Z}} &= (\widehat{z}_1,\dots,\widehat{z}_n)^\top, \\ \widehat{z}_i &= \widehat{\eta}_i + (Y_i - \widehat{\mu}_i)\,g'(\widehat{\mu}_i), \\ \widehat{\mu}_i &= g^{-1}(\widehat{\eta}_i), \\ \widehat{\eta}_i &= \mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}. \end{aligned} \tag{1.13}$$

This can be proved by multiplying equation (1.12) with $\left(\mathbf{X}^\top\widehat{\mathbf{W}}\mathbf{X}\right)$ and substituting values from (1.13), then we obtain

$$\left[\sum_{i=1}^{n} w(\widehat{\mu}_i)\mathbf{X}_i\mathbf{X}_i^\top\right]\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} w(\widehat{\mu}_i)\mathbf{X}_i\mathbf{X}_i^\top\right]\widehat{\boldsymbol{\beta}} + \sum_{i=1}^{n} w(\widehat{\mu}_i)g'(\widehat{\mu}_i)\,(Y_i - \widehat{\mu}_i)\,\mathbf{X}_i$$

which is equal to likelihood equations from (1.11), which we wanted to show.

Algorithm is based on iterations between $\widehat{\mathbf{W}}$ and $\widehat{\boldsymbol{\beta}}$, until the given precision is obtained. For a detailed description of IWLS method, see Dobson [2002].

1.5 Generalized Linear Mixed Model

1.5.1 Advantages of GLMM

In the previous section we described LMM and by using individual random effect, we were able to introduce the within-subject correlation. Furthermore, the "empirical Bayes" estimator for individual random effects was listed. Nevertheless, it still has a big limitation on the distribution of the response and range of the mean. Next, we presented GLM, which allows us to model random variables from exponential family of distribution and a more complex mean structure, but its disadvantage is that the observations of these variables are assumed to be independent. Due to these facts Generalized Linear Mixed Model (GLMM) was proposed. It combines all the benefits from LMM and GLM.

1.5.2 Definition of GLMM

GLMM is given by the following conditions

- We assume that Y_{it} follows unbalanced design with N individuals and n_i measures for each of them, like in LMM. Furthermore, independence between individuals is assumed as well, i.e., $\mathbf{Y}_i$ is independent with $\mathbf{Y}_j$ for $i \neq j$.
- Next, the random effects $\mathbf{b}_i$, $i = 1, \ldots, N$ are independent random vectors and $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0},\, \mathbf{D})$, where $\mathbf{D} \equiv \mathbf{D}(\psi)$ depends on parameter ψ.
- Given $\mathbf{b}_i$, components of $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ are conditionally independent, with density belonging to exponential family distribution
$$f(Y_{it}|\mathbf{b}_i) = \exp\left\{ \frac{Y_{it}\theta_{it} - b(\theta_{it})}{\varphi} + c(Y_{it},\, \varphi) \right\}.$$
- Then the conditional mean of Y_{it} given $\mathbf{b}_i$ is
$$\mu_{it} \equiv \mathsf{E}\,(Y_{it}|\mathbf{b}_i) = b'(\theta_{it})$$
and the conditional variance of Y_{it} given $\mathbf{b}_i$ has the following form
$$Var(Y_{it}|\mathbf{b}_i) = \varphi b''(\theta_{it}) \equiv \varphi V(\mu_{it}).$$
- Furthermore, it is assumed that μ_{it} is related to the linear predictor
$$\eta_{it} = \mathbf{X}_{it}^\top \beta + \mathbf{Z}_{it}^\top \boldsymbol{b}_i \tag{1.14}$$
through the link function $g(\mu_{it}) = \eta_{it}$.

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

1.5.3 Estimation of parameters

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

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\psi}|\mathbf{Y}) &= \prod_{i=1}^{N} f(\mathbf{Y}_i|\boldsymbol{\beta},\boldsymbol{\psi}) = \\ &= \prod_{i=1}^{N} \int_{\mathbb{R}^q} \prod_{t=1}^{n_i} f(Y_{it}|\boldsymbol{\beta},\mathbf{b}_i)\, f_b(\mathbf{b}_i|\boldsymbol{\psi})\, \mathrm{d}\mathbf{b}_i. \end{aligned} \tag{1.15}$$

Equation 1.15 can be expressed using canonical link $g(\cdot)$ and multivariate normal distribution f_b as follows

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

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

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

where C is constant with respect to $\boldsymbol{\beta},\boldsymbol{\psi}$ and $h(\cdot)$ is function of $\boldsymbol{\beta},\boldsymbol{\psi},\varphi,\mathbf{Y}$.

There are many approaches to maximize $\ell(\boldsymbol{\beta},\boldsymbol{\psi}|\mathbf{Y})$ and we will present a few of them. The first one is based on the Laplace approximation

$$\int_{\mathbb{R}^q}\exp\{Q(\mathbf{b})\}\,\mathrm{d}\mathbf{b} \approx (2\pi)^{-q/2}|-Q''(\tilde{\mathbf{b}})|^{-1/2}\exp\left\{Q(\tilde{\mathbf{b}})\right\}, \tag{1.16}$$

where $\tilde{\mathbf{b}}$ is the mode of $Q(\mathbf{b})$, i.e., $\tilde{\mathbf{b}} = \arg\max Q(\mathbf{b})$.

This approximation is obtained by replacing $Q(\cdot)$ in the integrand by the second order Taylor expansion of $Q(\cdot)$ around $\tilde{\mathbf{b}}$ and integrating the exponentiated quadratic function as a Gaussian density. Function $Q(\cdot)$ is equal to the sum of the second order Taylor expansion and remainder, which can be written as follows

$$Q(\mathbf{b}) = Q(\tilde{\mathbf{b}}) + Q'(\tilde{\mathbf{b}})(\mathbf{b}-\tilde{\mathbf{b}}) + \frac{1}{2}(\mathbf{b}-\tilde{\mathbf{b}})^\top Q''(\tilde{\mathbf{b}})(\mathbf{b}-\tilde{\mathbf{b}}) + R(\mathbf{b}),$$

where $R(\mathbf{b})$ is the remainder and can be expressed, e.g., Lagrange's form of the remainder, $R(\mathbf{b}) = (\mathbf{b}-\tilde{\mathbf{b}})^3 Q(c)'''/6$ for some c between $\mathbf{b}$ and $\tilde{\mathbf{b}}$. Consequently, integral can be re-written as

$$\int_{\mathbb{R}^q}\exp\{Q(\mathbf{b})\}\,\mathrm{d}\mathbf{b} \approx \int_{\mathbb{R}^q}\exp\left\{Q(\tilde{\mathbf{b}}) + \frac{1}{2}(\mathbf{b}-\tilde{\mathbf{b}})^\top Q''(\tilde{\mathbf{b}})(\mathbf{b}-\tilde{\mathbf{b}})\right\}\mathrm{d}\mathbf{b}$$

and after Gaussian integration, the expression 1.16 is obtained. For deeper theory, see 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 (1.15) 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 Fitzmaurice et al. [2004, Chapter 12.4] or 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 $\boldsymbol{\beta}$, $\mathbf{D}$, φ, the prediction for $\mathbf{b}_i$ can be calculated as follows

$$\widehat{\mathbf{b}}_i = \mathsf{E}\,(\mathbf{b}_i|\mathbf{Y}_i, \widehat{\boldsymbol{\beta}}, \widehat{\varphi}, \widehat{\mathbf{D}}),$$

which coincides with empirical “Bayes estimator” or BLUP for LMM as mentioned in Section 1.3. Such prediction is also not easy to obtain due to integration over the distribution of the unobserved random effects, $\mathbf{b}_i$, and again numerical methods must be used. Prediction of $\mathbf{b}_i$ 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. For further details see Fitzmaurice et al. [2004, Chapter 12.4]

1.6 Generalized Estimating Equations

In this section we present another approach called Generalized Estimating Equations (GEE) which is able to cope with correlated data within subjects. The main idea behind GEE is to generalize and extend the usual likelihood equations from GLM by including the covariance matrix of the vector $\mathbf{Y}$. The biggest advantage of this model is that we do not need to specify the whole distribution of the response. On the other hand, the mean structure, the mean-variance relationship and specification of the covariance structure need to be defined. The first two conditions are similar to GLM, see definition below.

Definition of GEE model

- Unbalanced design with independence between individuals $\mathbf{Y}_i$, $i = 1, \dots, N$ is assumed like in GLMM.

- Denote expected value of response $\mu_{it} \equiv \mathsf{E}\,(Y_{it})$, which depends on covariates, $\mathbf{X}_{it}$ as follows

$$g(\mu_{it}) = \eta_{it} = \mathbf{X}_{it}^{\top}\boldsymbol{\beta},$$

where $g(\cdot)$ is link function and together with the linear predictor η_{it} fully specify the mean structure μ_{it}.

- It is also assumed that the variance of each Y_{it} depends on the mean according to

$$Var(Y_{it}) = \varphi V(\mu_{it}), \tag{1.17}$$

where $V(\cdot)$ is a known variance function and $\varphi > 0$ is a scale or dispersion parameter, that can be known or may need to be estimated. It is worth mentioning that a GLS estimate is obtained by using identity link function.

- Furthermore, correlation between components of $\mathbf{Y}_i$ is represented by a *working correlation matrix* $\mathbf{C}_i \equiv \mathbf{C}_i(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is $s \times 1$ vector of unknown parameters.

The name “working” comes from the fact that the structure of $\mathbf{C}_i$ does not need to be correctly specified and asymptotic properties of estimate still hold. The corresponding working covariance matrix for i-th subject can be constructed as the product of standard deviations and working correlation matrix

$$\mathbf{V}_i = \varphi \mathbf{A}_i^{1/2} \mathbf{C}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2},$$

where $\mathbf{A}_i$ is diagonal matrix with $V(\mu_{it})$ along the diagonal.

As we mentioned, it does not need to specify the whole distribution, but due to form of variance from (1.17) we could consider that the (unknown) distribution belongs to the exponential family of distributions. However, it is not necessary.

1.6.1 Estimation of parameters

As we know GLS estimate of $\boldsymbol{\beta}$ from (1.6) minimizes the function

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

In GEE we have a similar situation, where estimator of $\boldsymbol{\beta}$ minimizes the objective function

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \boldsymbol{\mu}_i)^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i), \tag{1.18}$$

where $\mathbf{V}_i$ is treated as known and $\boldsymbol{\mu}_i \equiv (\mu_{i1}, \ldots, \mu_{in_i})^\top$ is vector with elements given by

$$\mu_{it} = g^{-1}(\mathbf{X}_{it}^\top \boldsymbol{\beta}).$$

Consequently, it can be shown that if a minimum of the function (1.18) exists, it must solve the generalized estimating equation

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i),$$

where $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta} \equiv \{\partial \mu_{it} / \partial \beta_k\}_{t,k=1}^{n_i,p}$ and $\mathbf{u}(\boldsymbol{\beta})$ is the so-called quasi-vector. The estimate of $\boldsymbol{\beta}$ solves equation $\mathbf{u}(\widehat{\boldsymbol{\beta}}) = 0$. Usually, parameters φ and $\boldsymbol{\alpha}$ from $\mathbf{V}_i$ are unknown, so they can be estimated by moment estimates. As in previous approaches, we also use the iterative algorithm. In one step, we estimate $\boldsymbol{\beta}$ and in the next step, we use this result for re-estimating $(\widehat{\varphi}, \widehat{\boldsymbol{\alpha}})$ until the required precision is obtained.

The most important properties of the estimate $\widehat{\boldsymbol{\beta}}$ are consistency, efficiency andasymptotic normality (for $N \to \infty$) with mean equal to $\boldsymbol{\beta}$, variance equal to $Cov\widehat{\boldsymbol{\beta}}$, which hold even when the working correlation matrix $\mathbf{C}_i(\boldsymbol{\alpha})$ is misspecified.

It is shown in Fitzmaurice et al. [2004, Chapter 11.3], that for large samples the variance of $\widehat{\boldsymbol{\beta}}$ can be expressed as follows

$$Cov(\widehat{\boldsymbol{\beta}})_S = \mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1},$$

where

$$\mathbf{B} = \sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i, \qquad \mathbf{M} = \sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} Cov(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i.$$

$\varphi, \boldsymbol{\alpha}$, $\boldsymbol{\beta}$ in $\mathbf{M}$ and $\mathbf{B}$ can be replaced by their estimates. Moreover $Cov(\mathbf{Y}_i)$ can be also replaced by $(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i)^\top$. Consequently, the estimate for variance of $\widehat{\boldsymbol{\beta}}$, known as the empirical or so-called sandwich estimator, is given by

$$\widehat{Cov}(\widehat{\boldsymbol{\beta}})_S = \\ = \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right)^{-1} \left\{ \sum_{i=1}^{N} \widehat{\mathbf{D}}_i^\top \widehat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i)^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right\} \left(\sum_{i=1}^{N} \widehat{\mathbf{D}}_i^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right)^{-1}. \tag{1.19}$$

The expression from (1.19) is consistent estimator of $Cov(\widehat{\boldsymbol{\beta}})$. If $\mathbf{V}_i$ is modelled correctly, $\mathbf{V}_i = Cov(\mathbf{Y}_i)$ and $Cov(\widehat{\boldsymbol{\beta}}) = \mathbf{B}^{-1}$.

In some cases, the sandwich estimator of $Cov(\widehat{\boldsymbol{\beta}})$ is not suitable, e.g., when the structure of data is strictly unbalanced or subjects cannot be grouped on the basis of having identical covariate design matrices. The same problem can be caused by a modest number of independent subjects (relative to the number of repeated measures). Therefore, a model-based $Cov(\widehat{\boldsymbol{\beta}})$ is more appropriate

$$Cov(\widehat{\boldsymbol{\beta}})_M = \mathbf{B}^{-1}, \quad \text{where} \quad \mathbf{B} = \sum_{i=1}^{N} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i, \tag{1.20}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and φ in $\mathbf{B}$ can be replaced by their estimates, which gives us the model based estimate of $Cov(\widehat{\boldsymbol{\beta}})_M$. However, in this case the choice of working covariance matrix $\mathbf{V}_i$ should be a close approximation of the true covariance $Cov(\mathbf{Y}_i)$, due to current variance structure of $\widehat{\boldsymbol{\beta}}$ from (1.20). Other properties of this estimate can be found in Liang and Zeger [1986] or in Fitzmaurice et al. [2004, Chapter 11.3].

1.6.2 Correlation structure

Despite the fact that asymptotic normality of $\widehat{\boldsymbol{\beta}}$ holds even when the correlation matrix is misspecified, a more precise choice of this matrix to the true one leads to more efficient estimates of $\boldsymbol{\beta}$. Parameter $\boldsymbol{\alpha}$ from $\mathbf{C}_i(\boldsymbol{\alpha})$ is assumed to be the same for all individuals and should be estimated in cases where it is unknown. Moreover, we have to specify how the correlation matrix should look like. There are several common choices for $\mathbf{C}_i(\boldsymbol{\alpha}) = \{c_{jk}\}_{j,k=1}^{n_i,n_i}$.

- The first and simplest one is *uncorrelated (or independent)* structure

$$c_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k. \end{cases}$$

- The opposite of the first one is *unstructured* correlation matrix

$$c_{jk} = \begin{cases} 1 & \text{if } j = k \\ \alpha_{jk} & \text{if } j \neq k. \end{cases}$$

- A mixture of the previous two is *exchangeable* structure

$$c_{jk} = \begin{cases} 1 & \text{if } j = k \\ \alpha & \text{if } j \neq k. \end{cases}$$

- Another choice is *m-dependent* correlation structure

$$c_{jk} = \begin{cases} 1 & \text{if } j = k \\ \alpha_{|j-k|} & \text{if } 0 < |j-k| \leq m, \\ 0 & \text{if } |j-k| > m. \end{cases}$$

- The last one is an $AR(1)$ correlation structure

$$c_{jk} = \alpha^{|j-k|}.$$

Choosing and estimation of covariance structure

In order to determine a suitable correlation structure and variance function, Pearson residuals must be defined as follows

$$r_{it} = \frac{Y_{it} - \widehat{\mu}_{it}}{\sqrt{V(\widehat{\mu}_{it})}}. \tag{1.21}$$

Now we describe a general strategy how to estimate parametrized correlations by the method of moments. Firstly, an estimate of $\boldsymbol{\beta}$ under working independence must be calculated. Furthermore, Pearson residual based on this model are expressed. Consequently, if the mean structure is correct, the following should hold

$$\begin{aligned} \mathsf{E}\, r_{it} &\approx 0 \\ Var(r_{it}) &\approx \varphi \\ \mathsf{E}\, r_{it} r_{ik} &\approx \varphi \left\{ \mathbf{C}_i \right\}_{tk}, \quad i = 1, \ldots, N, \quad t \neq k \in \{1, \ldots, n_i\}. \end{aligned} \tag{1.22}$$

Next, moment estimate of $\boldsymbol{\alpha}$ is calculated using these Pearson residuals, e.g., estimate of $\boldsymbol{\alpha}$ for 1-dependent correlation structure is given by

$$\widehat{\boldsymbol{\alpha}} = \frac{1}{\widehat{\varphi}} \frac{1}{\left(\sum_{i=1}^{N} n_i\right) - N - p} \sum_{i=1}^{N} \sum_{t=1}^{n_i - 1} r_{it} r_{it+1},$$

where p is number of components in vector $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\alpha}}$ is in this case a one dimensional estimate and $\widehat{\varphi}$ is a moment estimate of φ given by

$$\widehat{\varphi} = \frac{1}{\left(\sum_{i=1}^{N} n_i\right) - p} \sum_{i=1}^{N} \sum_{t=1}^{n_i} r_{it}^2.$$

1.6.3 Testing coefficients

This section deals with testing hypotheses for coefficients of vector $\boldsymbol{\beta}$. We assume that coefficient vector $\boldsymbol{\beta}$ consists of two sub vectors $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ with r and l components respectively, for which holds $p = r + l$. Thus, the vector of coefficient can be expressed as $\boldsymbol{\beta}^\top = (\boldsymbol{\gamma}^\top, \boldsymbol{\delta}^\top)$. The main aim of this section is testing hypothesis

$$H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0,$$

where $\boldsymbol{\gamma}_0$ is hypothesized value of $\boldsymbol{\gamma}$.

There are several approaches for constructing test statistics for hypothesis tests, e.g., likelihood ratio test, Wald test or score test. The first one mentioned cannot be applied to GEE directly because there is no likelihood underlying the model. However, this test can be used with modified assumptions, where the likelihood ratio is calculated under associated independence model. If zero hypothesis holds, then given test statistic has χ^2 distribution with r degrees of freedom.

The second mentioned test is calculated after model estimation. The test statistics are typically calculated without adjusting the degrees of freedom and use the sandwich estimate or model based estimate of $Cov(\widehat{\boldsymbol{\beta}})$ from (1.19) or (1.20). The generalized Wald test statistic with sandwich estimate of variance is given by

$$W = n(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\top \widehat{Cov}(\widehat{\boldsymbol{\beta}})_S^{-1} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0). \tag{1.23}$$

The test statistic is also assumed to follow χ^2 distribution with r degrees of freedom.

As we mentioned in the previous part dealing with the estimation of $Cov(\widehat{\boldsymbol{\beta}})$, the sandwich estimate is not always the best choice. If there are more covariates than subjects or panels, it can cause the sandwich estimate of variance to be singular. Due to this fact, an alternative to the generalized Wald test is the working Wald test, where the sandwich estimate of covariance from (1.23) is replaced by model based estimate of covariance matrix. To use this approach, it must be assumed that working correlation matrix $\mathbf{C}_i$ describes the true correlation structure of the data. Likelihood ratio test and score test are described in Hardin and Hilbe [2003, Chapter 4.5]

Model selection can also be done using modified information criteria, but it should be used only in cases, where there is no other way to choose between models. Several information criteria for GEE models are presented in Hudecová and Pešta [2013].

2. Introduction to reserving theory

This chapter deals with claims reserving, which is the main problem in non-life insurance. Models for life insurance are rather different due to the structure of products, nature of claims, risk drivers, term of contracts etc. These are the reasons for the separation of life and 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 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 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.

There are two main types of reserves. The first one is reserves for claims that have been reported but have not been settled yet, so called RBNS (Reported But Not Settled). The second one is reserves for claims that have occurred but have not been reported, so called IBNR (Incurred But Not Reported). The last mentioned often contains reserves for not enough reported incurred claims IBNeR (Incurred But Not enough Reported).

It is worth mentioning that claims costs are often impacted by inflation. The main effect of inflation is not related on the salary or price but on 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.

2.1 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, so-called claims development triangle is used, see Table 2.1. Let Y_{it} stand 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 are placed in right-angled isosceles triangle $\{Y_{it}\}$, where $i = 1,\ldots,n$ and $t = 1,\ldots,n+1-i$.

Let us denote a random variables C_{it}, cumulative payments or cumulative claims,

Accident	Development year t						
year i	1	2	$\cdots$	t	$\cdots$	n-1	n
1	Y_{11}	Y_{12}	$\cdots$	Y_{1t}	$\cdots$	Y_{1n-1}	Y_{1n}
2	Y_{21}	Y_{22}	$\cdots$	X_{2t}	$\cdots$	Y_{2n-1}	
$\vdots$	$\vdots$	$\ldots$	$\ddots$	$\ldots$	$\ldots$		
i	Y_{i1}	$\cdots$	$\cdots$	Y_{it}			
$\vdots$	$\vdots$	$\ldots$	$\ldots$				
n	Y_{n1}						

Table 2.1: Run-off triangle for incremental claim amounts X_{it}.

in origin year i after t development years, e.i. $C_{it} = \sum_{k=1}^{t} Y_{ik}$. Observations of C_{it} for $i+t-1 \leq n$ form a cumulative run-off triangle. All our effort is concentrated on estimating the ultimate claims amount C_{in} and consequently, on calculating reserves for all accident years $i = 2, \ldots, n$ as follows

$$R_i^{(n)} = C_{in} - C_{in+1-i}. \tag{2.1}$$

This text deals only with reserves defined in 2.1 and does not assume any tail factor.

2.2 Basic reserving methods

Early methods for distributing risk were practiced by Chinese and Babylonian traders as long ago as the 3rd and 2nd millennia BC, respectively. Modern insurance began in Europe where it became far more sophisticated and specialized. Insurance as we know it today is dated to 1667 and was founded after the Great Fire of London in 1666.

Claims reserving has significantly developed relatively recently. The first deterministic reserving model, original chain-ladder, was developed by Fisher & Lange in 1973. Later, more complex approaches were needed, so a random part was added to the existing models, which resulted in stochastic models. The basis for these models was founded by Mack in 1993 and was built on assumption of proportionality of columns in a run-off triangle. These stochastic approaches are presented and described in Wüthrich and Merz [2008] or England and Verrall [2002]. Almost all of these proposed stochastic models require independence of incremental claims Y_{it}, which in practice does not often hold. Due to this, models such as GLMM and GEE were introduced because of their ability to cope with dependencies within subjects.

3. Claims reserving in panel data framework

Previous chapters described all necessary theory in general. In this chapter, theory of GLMM and GEE from Chapter 1 are applied on claims reserving presented in Chapter 2. The advantages of these models seem to be a suitable solution for the problem which is possible dependence among the incremental claims within accident year i. This approach is also pointed out in a paper written by Antonio and Beirlant [2007].

Notation and terminology from Chapter 2 are used in GLMM and GEE models for claims reserving. All general framework from Chapter 1 is adjusted to fit data structure of claims.

The first section describes claims reserving in the GLMM framework. Discussion is focused on a suitable choice of linear predictor, link function and distribution of dependent variable.

Application of GEE to claims reserving is presented in Section 3.2. Just as for GLMM, a suitable linear predictor is discussed as well as link function. Next, the choice of working correlation matrix and variance function is described. Finally, testing of coefficients and their reasons are discussed.

Prepared GLMM or GEE framework is applied on incremental claims Y_{it} from run-off triangle in Table 2.1, which represents known observations of random variables Y_{it}. The lower right part of the rectangle, $Y_{it},\ i = 1,\dots,n,\ n \geq t > n-i+1$, is unknown and needs to be predicted in order to estimate the amount of total reserves which equals to

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

Sections 1.5, 1.6 deal with GLMM and GEE approaches using unbalanced design, i.e., $t = 1,\dots n_i$, and $i = 1,\dots, N$, where n_i stands for number of observations in i-th subject. Rewritten into reserving data structure $n_i = n - i + 1$, where n is current year and i is accident year, which symbolizes i-th subjects in GLMM or GEE. From now on, it holds that subject or accident year i is from the range $1,\dots, n$.

3.1 GLMM method for claims reserving

3.1.1 Link function

This section describes the process of finding a suitable choice of GLMM. First of all, link function must be specified. 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 mentioned facts, commonly used link functions are log-link, $g(\cdot) = \log(\cdot)$, or identity link function, $g(\cdot) = (\cdot)$. By using the log-link, the response in run-off triangle is assumed to be positive. If a few negative values occur, log-link can be applied, but non positive values must be replaced by positive ones close to zero. However, 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. However, we will only use it on datasets with positive incremental data in run-off triangle.

3.1.2 Linear predictor

Next, the choice of the linear predictor is important for the model. Due to the very specific structure of our data, there are not many choices for the linear predictor from 1.14. First, the simple one can be written as

$$\eta_{it} = \beta_0 + b_i + \beta_t, \tag{3.1}$$

where β_0 is intercept, same for all accident years i. In order to avoid over-parametrization, 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. 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

$$\eta_{it} = \beta_0 + b_{i0} + \beta_t + b_{it}, \tag{3.2}$$

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 (3.2) can be written also as follows

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

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

$$\mathbf{X}_{it} = (1, d_{2t}, \ldots, d_{nt})^\top.$$

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

The practical part of this thesis deals only with the simplest model of linear predictor 3.1, because the aim of this thesis is to compare the suitability of GEE and GLMM methods in a certain way. In order to do this, a similar linear predictor must be chosen for both methods, so for this purpose only random intercept is chosen, which can be comparable in some sense to the fixed coefficients for accident years in GEE.

3.1.3 Distribution of incremental claims

We assume conditional distribution of incremental claims Y_{it} given $\mathbf{b}_i$ belongs to exponential family distribution. As it was mentioned before, significant members of exponential family distribution are Gaussian, inverse Gaussian, Poison, over-dispersion Poison and gamma distribution. Poison distributions are not considered because they are used for modelling non negative integers and amount of claims are not supposed to be integers at all.

The choice of a suitable distribution function is made according to precision of the fitted values and residual diagnostic, where the relation between mean and variance is investigated. Connection between mean and variance is described through variance function $V(\cdot)$ and for chosen distributions it is given by

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

In cases, where we really could not decide on a suitable model according to residual diagnostics or precision of fitted values, information criteria introduced in Bolker et al. [2009] can be used. However, we try to avoid using such criteria because there is enough information in the residual diagnostic for this purpose.

3.2 GEE method for claims reserving

Let's go on to the application of GEE models to claims reserving. As it was already mentioned, the advantage of this method in contrast to GLMM is that the distribution of claims does not need to be specified. However, it brings other issues that need to be dealt with, i.e., specification of variance function and working correlation matrix. GLMM, as well as GEE, is applied on incremental claims Y_{it}.

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.

3.2.1 Linear predictor

Mean structure in GEE is different than in 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

$$\eta_{it} = \gamma + \alpha_i + \beta_t, \tag{3.3}$$

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

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

where $\boldsymbol{\beta} = (\gamma, \alpha_2, \ldots, \alpha_n, \beta_2, \ldots, \beta_n)^\top$ and vector $\mathbf{X}_{it}$ is defined using dummy variables as follows $(1, d_{2i}, \ldots, d_{n,i}, d_{2t}, \ldots, d_{nt})^\top$.

There are several other linear predictors, e.g., Hoerl curve with the log-link function, which can be parametrized by vectors $\mathbf{X}_{it}$ and $\boldsymbol{\beta}$ as follows

$$\mathbf{X}_{it} = (1, d_{2i}, \ldots, d_{n,i}, 2 \times d_{2t}, \ldots, n \times d_{nt}, d_{2t} \times \log 2, \ldots, d_{nt} \times \log n)^\top,$$

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

which leads to linear predictor

$$\eta_{it} = \log(\mu_{it}) = \gamma + \alpha_i + t\beta_t + \lambda_j \log j.$$

However, we should realize that we have only $n(n+1)/2$ observations and model with $3(n-1)+1$ parameters, what 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 3.3 is used in the practical part, even though it still has a lot parameters.

The impact of inflation on the development of claims was discussed in Chapter 2, but in cases where the market is stable and inflation does not effect the amount of claims much, more simpler models should be taken into account, e.g., models with fewer numbers or even without accident year factors. Such models lead to better interpretations, the estimates become more precise and efficient, which are very important and practical properties.

Due to this, testing coefficient of particular accident or development year could be made in order to obtain suitable model with appropriate number of coefficients. In order to do this, the Wald test described in Section 1.6 can be used.

To sum up, our purpose is not to find the most complex model, which is difficult to interpret, but to find a model which is reasonable and provides suitable properties.

3.2.2 Variance function

The next thing that needs to be specified is the variance function. It is possible to define various functions but, in order to avoid confusion in the amount of models used in the practical part, only three basic variance functions are assumed

$$V(\mu_{it}) = \begin{cases} 1, \\ \mu_{it}, \\ \mu_{it}^2. \end{cases}$$

It can be seen that two of them are the same as for GLMM, i.e., Gaussian, Gamma models for which variance functions equal 1 and μ_{it}^2, respectively.

3.2.3 Correlation structure

Finally, determination of working correlation structure is described in this part. In Section 1.6 several correlation structures were introduced. However, only a few of them are used in our practical analysis, due to the data structure of claims and the number of parameters that need to be estimated in a working correlation matrix. The choice is reduced to independent, exchangeable and $AR(1)$ structure, where one or no parameter needs to be estimated. The unstructured and m-dependent structure are not considered due to the high number of parameters.

The appropriateness of a model with given correlation structure is made after fitting the model, according to properties of Pearson residuals 1.22, as well as the whole residual diagnostic. In the practical part, we will use plot of fitted values with respect to observed values to see how good the model will fit the data. Next, QQ plot, scatter plot, histogram of residuals and plot of classical residuals will be listed as well. These criteria are used to assess GLMM as well.

As we already mentioned in Section1.6, when we are not able to decide between GEE models, we can use information criteria, which can help us decide.

4. Practical application of models

As it was already pointed out, this practical part focuses on models with log-link function and linear predictors in form 3.1 for GLMM and 3.3 for GEE. There are three proposed models for GLMM, first one with Gaussian, second one with inverse Gaussian and the last one with gamma distribution.

For GEE, nine models are analysed. They differ by the choice of variance function $V(\mu_{it})$ which can be equal to one, μ_{it} or μ_{it}^2. Furthermore, the model is defined by correlation structure which, in our case, can be independent, exchangeable or $AR(1)$. It should be noted that random effects for accident years will be predicted as well.

For the purpose of clarity in our tables and figures, abbreviations for GEE models are introduced. First letters stands for correlation structure, i.e., AR for $AR(1)$, IND for independent and EX for exchangeable correlation structure. Letters behind the underscore denote variance function, i.e., 1 for $V(\mu_{it}) = 1$, L for $V(\mu_{it}) = \mu_{it}$ and Q for $V(\mu_{it}) = \mu_{it}^2$. For example, GEE model with exchangeable correlation structure and variance function equal to 1 is label EX_1.

Everything is now prepared to apply theory to real data.

4.1 Datasets

The whole chapter deals with datasets of claims from the National Association of Insurance Commissioners (NAIC) database, which can be found in Meyers and Shi [2011].

The database contains cleaned claims developments of three lines for business (Private passenger auto liability/medical, Commercial auto/truck liability/medical, Workers' compensation) for all 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 then to test its performance. Then, a retrospective analysis using all data including lower triangle is made as well.

First of all, datasets populated with cumulative claims are transformed to incremental ones and then the datasets, which contain non positive values in the upper incremental triangle are excluded, due to the log-link function, which we decided to use in Chapter 3.

Next, data analysis is performed on 16 datasets from Private passenger auto liability/medical (146), 12 datasets from Commercial auto/truck liability/medical (158) and 30 datasets from Workers' compensation (132). The total number of datasets including those where non positive incremental values occurs is listed in parenthesis.

Three GLMM and nine GEE models are fitted on all of these datasets. Then, residual diagnostics on the upper triangles are performed in order to pick the best model without knowing the lower triangle as would be the case in reality. Next, a residual diagnostic is made using the whole rectangle and real reserves are computed, in order to check the change of the residual diagnostic and the precision of prediction respectively. The aim of this analysis is to try to find a suitable dataset for the GLMM approach when the GEE method is not proper, then to find a dataset where GEE fits better than GLMM and last, to find a dataset where all these methods do not work well.

The best outcome would be, when we could generally say that in case when our data has certain behaviour then GLMM is more suitable or vice versa.

According to the facts mentioned above, in following sections, three main and one additional datasets are presented on which the advantages of the individual models are described. Diagnostic figures are shown only for selected models because otherwise it would be 12 diagnostic figures for each dataset, what would cause only opacity in text. Nevertheless, all these figures could be generated by R scripts which are attached in Appendix A. Software R is chosen for practical analysis because it offers a wide range of packages that are useful for our purpose.

4.2 Claims reserving within GLMM

4.2.1 Dataset

The first dataset, Hastings Mut. Ins. Co., from line of business Workers' compensation, is chosen in order to show why in some cases, GLMM can be better than GEE. It can be seen from incremental claims in Table 4.1, that the time-lag between the claims occurrence and the claims payments in this dataset is pretty huge. The main amount of claims is paid in the second development year, not in the first one, as is usual. This property is better illustrated in Figure 4.1, where developments for each accident year are shown.

Accident	Development year j									
year i	1	2	3	4	5	6	7	8	9	10
1	1117	1116	630	304	369	80	152	46	37	24
2	1603	2007	1232	1052	432	259	188	109	32	
3	2136	2600	2134	1670	562	340	147	86		
4	2706	3009	2026	928	426	256	83			
5	3229	4582	2757	1229	546	331				
6	3381	4793	2524	1404	841					
7	3969	4529	2897	1148						
8	3661	4023	2828							
9	3687	3848								
10	3406									

Table 4.1: Observed run-off triangle for incremental payments.

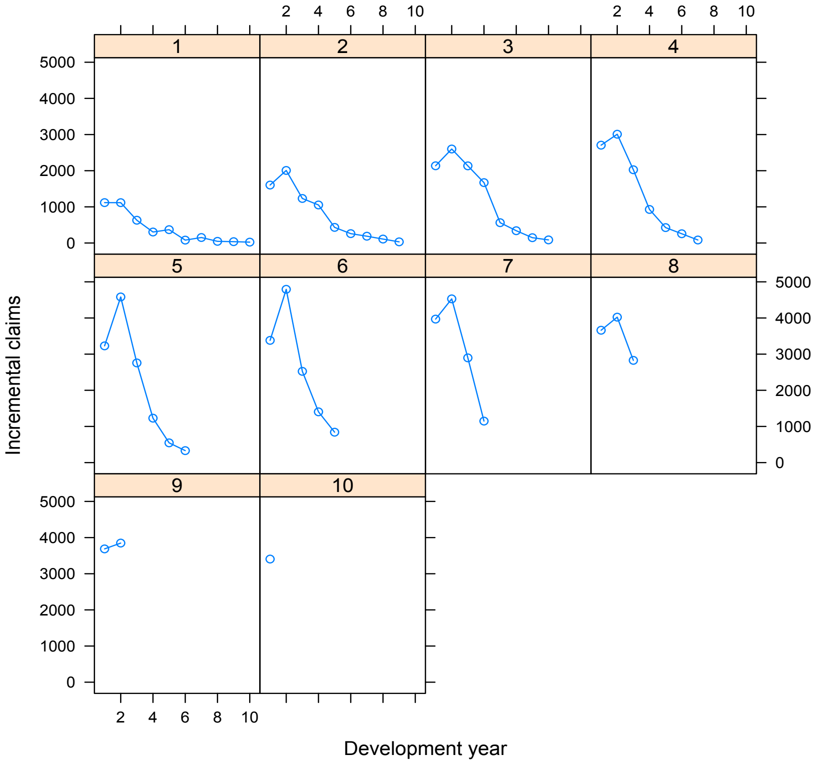

Figure 4.1: Claims development for each accident year.

After a more detailed analysis of the data, it can be seen that the size of the peaks varies a lot 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. This is a very important property of the data and should be taken into account. This could be a case where GLMM fit better, due to Bayesian approach used for prediction of random accident year factors as mentioned in Chapter 1. This Bayesian property could handle such variation of the data properly. Similar behaviour appears in more analysed datasets, but this is the most representative one. However, it does not mean that after such variation occurs, only GLMM should be chosen. Nevertheless, we should be aware of this during the model selection.

Next, we present the lower part of the rectangle in Table 4.2, which will be used later for a comparison with the predicted values of incremental payments and also for a retrospective residual diagnostic. From rectangle of incremental claims, it can also be seen that the highest amount of the last accident year claims is paid in the second development year as well. Nevertheless, there is not such significant peak as in other accident years, which is an effect of the above mentioned behaviour of the data and may have an impact on the accuracy of the predicted reserves.

Accident	Development year j									
year i	1	2	3	4	5	6	7	8	9	10
1										
2										29
3									46	22
4								49	39	42
5							264	209	75	88
6						442	200	184	85	81
7					481	312	196	69	72	56
8				1065	488	249	235	73	62	50
9			2137	1083	589	411	317	122	171	52
10		3882	1474	832	511	195	135	149	106	46

Table 4.2: Lower triangle of incremental payments.

4.2.2 Residual diagnostic

Firstly, all selected models from Chapter 3 are fitted on the upper triangle and then residual diagnostic based on the upper triangle is generated as well. Next, comparison of all GLMM and GEE models is made. Due to the residual diagnostic in Figure 4.2, the GLMM Gaussian model was chosen.

The first diagnostic plot in Figure 4.2 is a QQ plot, where points are expected to be placed near the line, nevertheless, we do not have sufficient number of observations for such asymptotic behaviour, thus diagnostic is not very significant for this purpose. The next one is a plot of fitted values with respect to their observed values, where all points should be and are placed near the diagonal, which implies that Gaussian model fits the data well and could be a suitable choice. According to the scatter plot of residuals, no correlation between residuals from development year t and $t-1$ in individual accident years is observed, due to the horizontal dashed line, which is the result of basic linear regression applied on the plotted points. This fact is in line with our assumption of the conditional independence.

Skewness of the histogram may be caused by insufficient number of observation as we mentioned by the QQ plot. Next to the histogram, a plot of Pearson residuals with respect to fitted values is listed. There is no visible pattern, i.e., variance is not increasing variance with higher fitted values, which coincides with our expectation. The last plot illustrates a classical residual with respect to t which goes through all rows of run-off triangle. Residuals are symmetrically placed around zero as we expected.

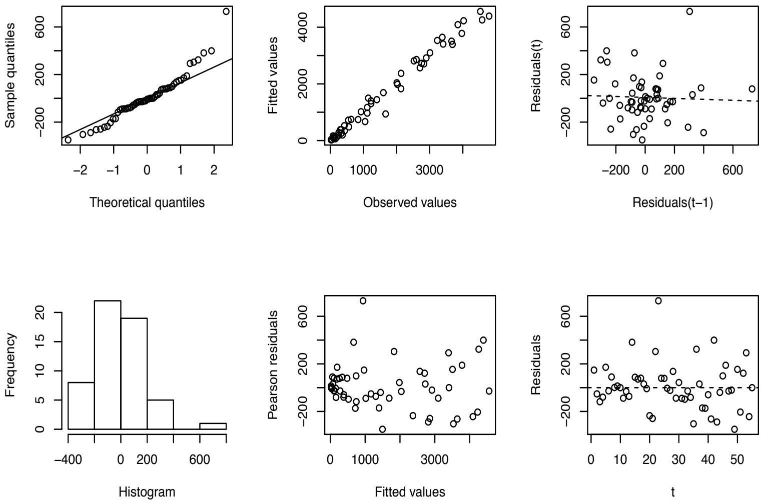

Figure 4.2: Residual diagnostic generated using upper triangle.

Based on our previous diagnostic and the mentioned behaviour of the data, Gaussian model was chosen as the final one. Here, it is time to look at residual diagnostic in Figure 4.3 based on the whole rectangle. The QQ plot gets better, but still not as good as we expected. Placement of the points implies heavy tails, which are visible in the histogram as well. The histogram looks more like Gaussian one but still a little bit skewed. The next plot, Pearson residuals with respect to fitted values, is still in line with our assumption except for a few outliers. The last plot shows classical residuals, which are uniformly placed around zero as it was in Figure 4.2.

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

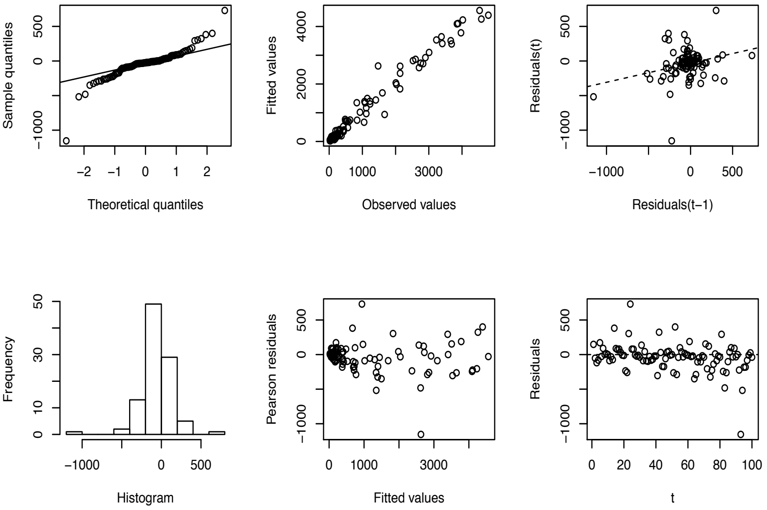

Figure 4.3: Residual diagnostic generated using all rectangle.

4.2.3 Claims reserving

Figure 4.4 illustrates all GLMM predictions for each accident year beginning directly after the vertical red line. Fitted values are graphed before the red line. Based on the fitted values, gamma model seems to be suitable as well. Residual diagnostic for this model is similar to the Gaussian model, except for a plot of Pearson residuals with respect to fitted values in Figure 4.5, which implies that our GLMM with gamma distribution is not suitable for our data. Let's focus more on predicted values. It is hard to say from Figure 4.4, which GLMM model has the most precise over all prediction.

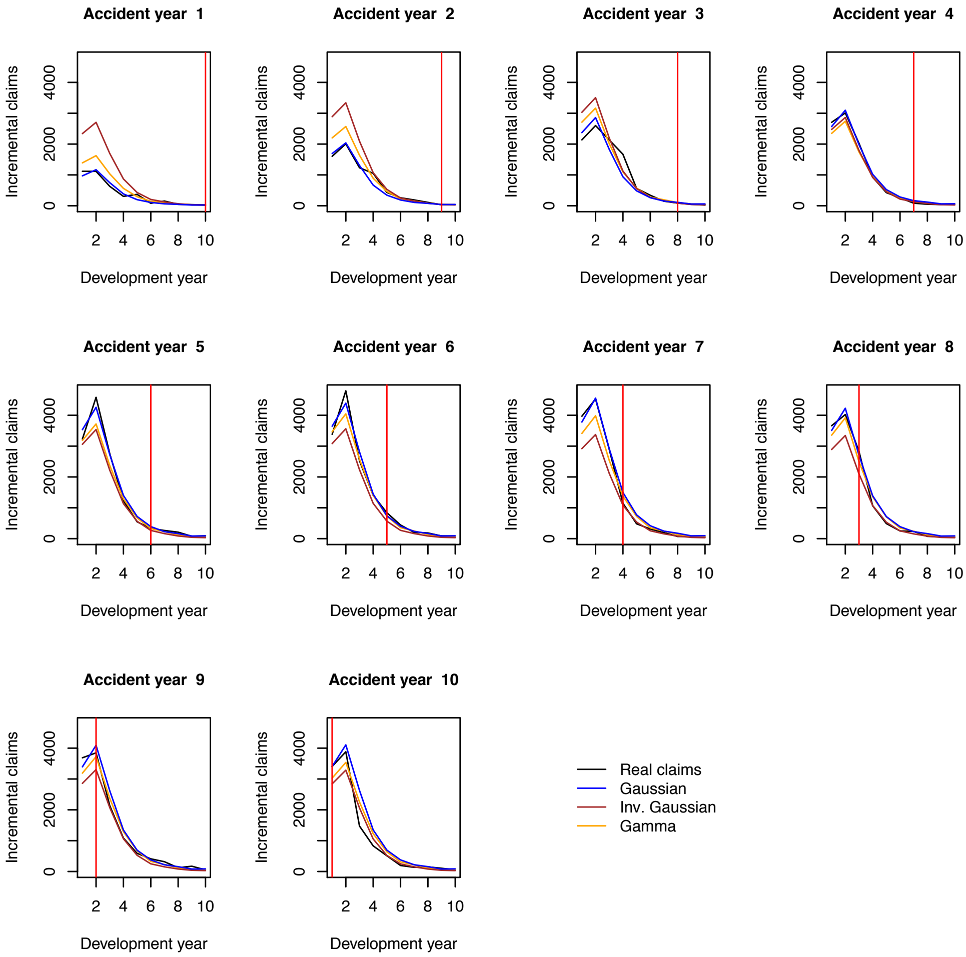

Figure 4.4: Fitted and predicted values vs. real values.

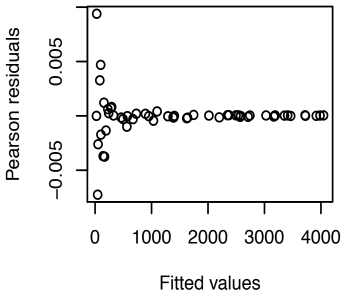

Figure 4.5: Pearson residuals vs. fitted values generated using upper triangle.

This fact is clearer in Table 4.3, where, using information from the lower triangle, real reserves are computed and compared with all GLMM and GEE predicted reserves. Moreover, the Mack-chain-ladder model is used for this purpose as well. However, we should be aware of the fact that this model has different, weaker, assumptions and subsequently different interpretation. This can cause that it has worse prediction, but on the other hand, in cases where assumptions of more restrictive models do not hold it may still provide reasonable results. Table 4.3 shows that reserves predicted by the inverse Gaussian model are the closest to the real one despite the worst residual diagnostic. Almost all GEE predictions of reserves are much higher except the model with independent correlation structure and variance function equal to one, which has the same prediction of reserves as the Gaussian model in GLMM, but still is not better than the gamma or inverse Gaussian model. It is worth to mentioning that prediction of reserves using the Mack-chain-ladder model is the same as for GEE model with independent correlation structure and linear variance function, what can be also seen in following datasets. To sum up, according to precision of the prediction and residual diagnostic GLMM approach is more suitable for this dataset.

Reserves

Real	Mack	GLMM	Predictions	GEE models	Predictions
17475	22625	Gaussian	22033	*AR_Q*	23028
		Inv. Gaussian	16077	*AR_L*	22569
		Gamma	19672	*AR_1*	22145
				IND_Q	22659
				IND_L	22625
				IND_1	22033
				EX_Q	22659
				EX_L	23176
				EX_1	25196

Table 4.3: Real and predicted reserves.

As it was mentioned in Chapter 3, our GLMM use basic linear predictor 3.1 with random intercept and logarithmic link function. Table 4.4 lists estimated coefficients for development years and predicted random factors for accident years.

Coefficient $\widehat{\beta_2}$ is positive as was expected, due to the time lag that occurred in the second year. All other estimated development factors are negative and decreasing except for development year 10, where there is a slight increase of estimated coefficient in comparison with previous one. Predicted random factors fluctuate around the intercept $\widehat{\beta_0} = 7.896$.

Accident year i	Predicted $\widehat{\beta_0} + \widehat{b_i}$	Development year t	Estimation $\widehat{\beta_t}$
1	6.881	1	0
2	7.430	2	0.187
3	7.773	3	-0.259
4	7.852	4	-0.926
5	8.170	5	-1.591
6	8.203	6	-2.203
7	8.241	7	-2.747
8	8.164	8	-3.081
9	8.130	9	-3.743
10	8.132	10	-3.698

Table 4.4: Prediction and estimation of coefficient from linear predictor.

To sum up, if differences between the first and second development year vary a lot through the accident years without any "trend" occurring and according to residual diagnostic, some of the GLMM seem to be suitable, then we should really take them into account. However, if the GEE model clearly has a better diagnostic, then this model should be our choice, despite the variation of the data. This will be presented in the following section.

4.3 Claims reserving within GEE

4.3.1 Dataset

The strengths of GEE models are shown on dataset Millers Mut Ins. Assoc. from Workers' compensation. Just as with the first dataset, this one also contains the lower triangle which is further used for calculation of real reserves and retrospective residual diagnostic.

The process of finding a suitable model is the same as for the first dataset. However, when the residual diagnostic implies that GEE models are the more suitable ones and some of them have similar good results, then special attention is paid to the mentioned properties of Pearson residuals. Furthermore, in order to optimize the number of coefficients, testing of their significance is performed.

So back to the dataset description. Table 4.5 presents run-off triangle of incremental payments and is graphically interpreted in Figure 4.6. As in the previous dataset, similar behaviour of peaks without any "trend" is observed. We should be aware of this fact during the model selection.

Accident	Development year j									
year i	1	2	3	4	5	6	7	8	9	10
1	1769	3186	1457	1010	476	256	76	46	39	42
2	2605	3011	1713	1141	86	561	354	244	306	
3	2277	3407	1580	1012	544	188	202	125		
4	2062	2862	1357	811	417	298	159			
5	1914	2534	1408	663	248	225				
6	1737	2878	1382	754	450					
7	1959	2263	1057	542						
8	1381	1869	938							
9	1565	2025								
10	1475									

Table 4.5: Observed run-off triangle for incremental payments.

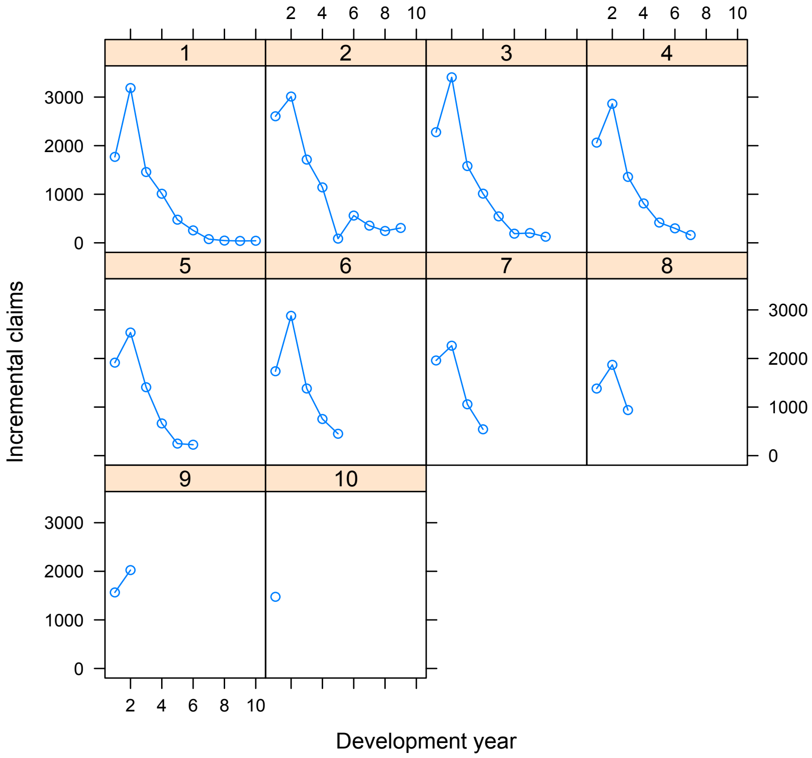

Figure 4.6: Claims development for each accident year.

The same behaviour of the data, as in the first example, is observed by adding the lower triangle from Table 4.6, where the last accident year value in first development year (upper triangle) is a bit higher than the value in second development year (lower triangle). This development differs from upper triangle pattern and consequently, has an impact on the amount of real reserves as well.

Accident	Development year j									
year i	1	2	3	4	5	6	7	8	9	10
1										
2										179
3									102	38
4								126	122	113
5							189	102	23	22
6						195	-77	91	44	31
7					333	201	140	49	53	108
8				497	179	109	-2	47	26	58
9			864	543	336	129	156	86	77	0
10		1449	827	319	419	342	213	140	130	131

Table 4.6: Lower triangle of incremental payments.

4.3.2 Residual diagnostic

Let's go on to the model selection. According to the residual diagnostic of all models, GEE models fit the data much better than GLMM. Especially models with exchangeable and $AR(1)$ correlation structure, with the same variance function equal to one have quite good results in diagnostic figures. In order to determine which correlation structure is more appropriate, we focus on the third property of Pearson residuals from 1.22. So we plotted these residuals from the model with independent correlation structure and variance function equal to one in Figure 4.7, where red points symbolize arithmetic means for given values. Based on this figure, we are not able to choose one of these two models due to a lot of outliers which influence these arithmetic means and an insufficient amount of data. Finally, we decide for model with exchangeable correlation structure because, as it will be shown later, some coefficients are not statistically significant and simpler linear predictor can be used. Unlike the model with $AR(1)$ correlation structure where all coefficients are statistically significant.

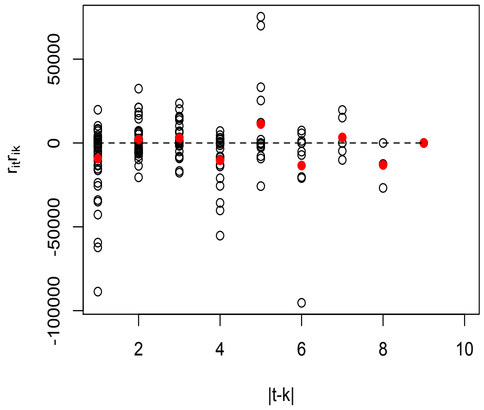

Figure 4.7: Products of Pearson residuals with respect to their distance within accident year based on the upper triangle.

Figure 4.8 shows residual diagnostic for chosen model with exchangeable correlation structure, where the first diagnostic is a QQ plot, where all points are placed on the line except some outliers. Then, a plot of observed values with respect to fitted values is listed, where all points lay almost on a diagonal line as is expected. Next, a scatter plot of residuals indicates a very small correlation between residuals in development year t and $t-1$. The histogram is a bit skewed but close to Gaussian one. In the plot of Pearson residuals with respect to fitted values, no pattern is visible which is a very important indicator for the right choice of variance function as well. The last plot of classical residuals looks like we expected due to their symmetrical placement around the zero. So, according to all above mentioned, the model with exchangeable correlation structure and variance function equal to one seems to be the most suitable.

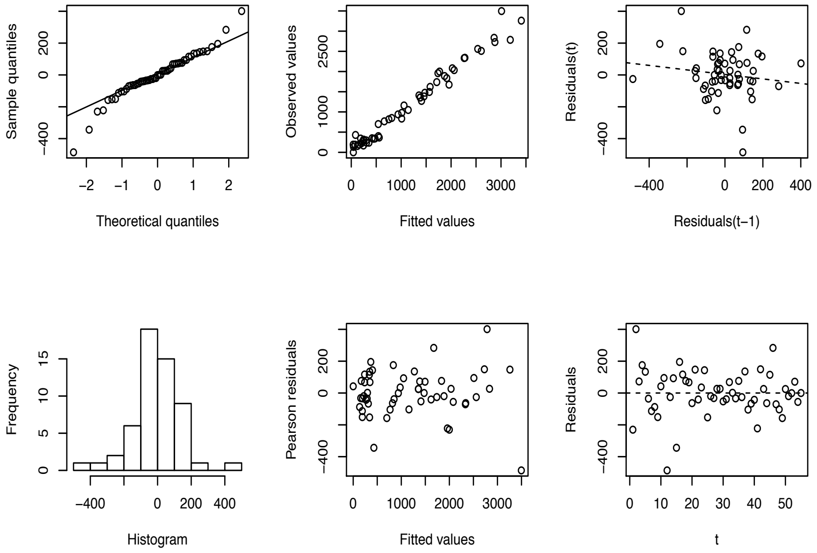

Figure 4.8: Residual diagnostic generated using upper triangle for model EX_1.

In the following section, information from the lower triangle is used for the purpose of reserve calculation and retrospective residual diagnostic in Figure 4.9. It is worth mentioning that the change of the scatter plot, where the almost horizontal dashed line does not imply linear dependency between residuals and residuals from previous development year. The histogram changed a bit as well, i.e., looks more like Gaussian one. Other indicators still look pretty good and have no significant changes. To sum up, it confirms the suitability of the chosen model.

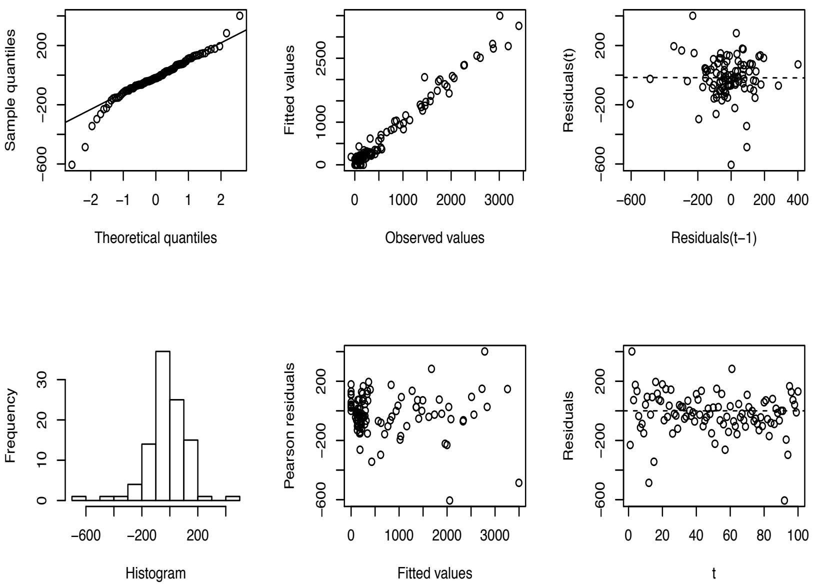

Figure 4.9: Residual diagnostic generated using all rectangle.

4.3.3 Claims reserving

In order to see the reasons for the gap between predicted reserves and the real one, the development of real claims, fitted values and predictions of GEE models with exchangeable correlation structure are illustrated in Figure 4.10. The fitted values are pretty close to the real amounts of incremental claims except for the large difference in the second development year of accident year 2. It is also visible that our final model is the closest to the real value in the mentioned point.

Let's focus more on the predicted values which are quite precise, except for the last accident year, where the biggest gap occurs in the second development year. This fact causes the predicted reserves to be quite higher than the real one.

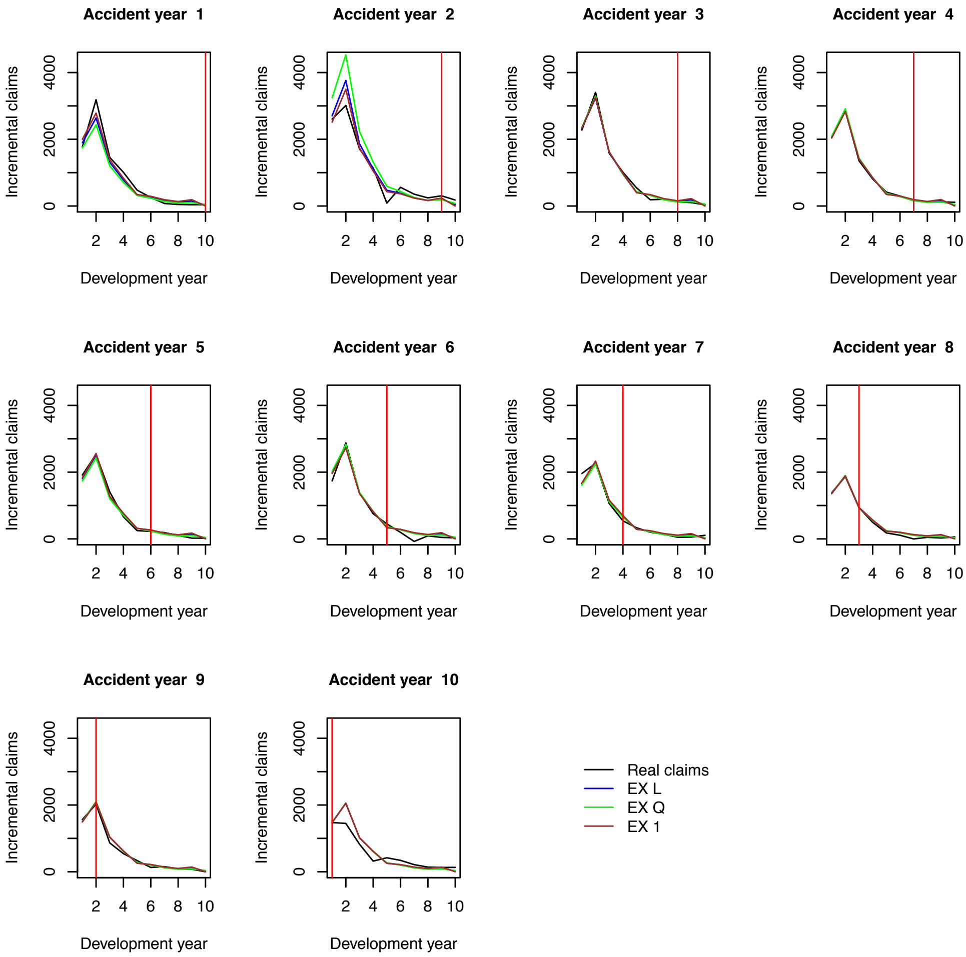

Figure 4.10: Fitted and predicted values vs. real values.

Data from Table 4.6 is finally used for the computation of real reserves in Table 4.7, where predictions of reserves for all models are presented. Table 4.7 shows that all GEE models predict real reserves much better than the GLMM ones.

Our final model is the second worst in accuracy of prediction from all GEE models, but still much better than all GLMM. However, GEE model with $AR(1)$ correlation structure and variance function equal to one, which also has similar reasonable residual diagnostic has the third best prediction of total reserve. In this case, the Mack-chain-ladder model provides reasonable prediction as well.

Reserves					
Real	Mack	GLMM	Predictions	GEE models	Predictions
9259	11064	Gaussian	14857	*AR_Q*	10817
		Inv. Gaussian	14190	*AR_L*	11084
		Gamma	13143	*AR_1*	10953
				IND_Q	10656
				IND_L	11064
				IND_1	11194
				EX_Q	10656
				EX_L	10994
				EX_1	11171

Table 4.7: Real and predicted reserves.

4.3.4 Testing coefficient

Now, testing of coefficients for our final GEE model is discussed. The advantages of a simpler model in the sense of the number of coefficients was described in Chapter 3. In order to do this a Wald test is performed. Chosen software provides the generalized Wald test with sandwich estimate of variance as mentioned in Chapter 1.

In order to avoid misinterpreting the results, the following must be explained. Notation $\widehat{\alpha}_i$ $(\widehat{\gamma})$ from Tables 4.8 and 4.9 means that in the first row, i.e., accident year $i = 1$, estimate $\widehat{\gamma}$ is listed otherwise $\widehat{\alpha}_i$. According to the results in Table 4.8, coefficients for accident years 4 and 6 are not statistically significant on given level equal to 5 %. This fact is quite reasonable because the values of these coefficient are very close to zero.

Accident	Estimation	P-values	Development	Estimation	P-values
year i	$\widehat{\alpha}_i$ $(\widehat{\gamma})$	Wald test	year t	$\widehat{\beta}_t$	Wald test
1	7.60	$\doteq 0$	2	0.331	$\doteq 0$
2	0.228	$\doteq 0$	3	-0.368	$\doteq 0$
3	0.158	$\doteq 0$	4	-0.874	$\doteq 0$
4	0.0180	0.518	5	-1.765	$\doteq 0$
5	-0.0842	0.011	6	-1.926	$\doteq 0$
6	-0.0203	0.589	7	-2.357	$\doteq 0$
7	-0.177	$\doteq 0$	8	-2.703	$\doteq 0$
8	-0.388	$\doteq 0$	9	-2.352	$\doteq 0$
9	-0.291	$\doteq 0$	10	$-608*10^{16}$	$\doteq 0$
10	-0.304	$\doteq 0$			

Table 4.8: Estimation of coefficient from linear predictor with corresponding P-values.

It is not suitable to put all coefficients for these two accident years equal to 0 based on current results. It would be possible to use the test of join hypotheses, but in this analysis, we apply a backward elimination approach using P-values as a criterion. Hence, the coefficient for accident year 6 with the highest P-value is excluded from the model and then again whole residual diagnostic is made in order to ensure that the model still fits the data well. In this case, the residual diagnostic does not change significantly.

Next, a test of coefficients is computed on the simpler model without effect for accident year 6. According to the results of this test, coefficient for accident year 4 has the highest P-value (0.949), so this factor is excluded from the model as well. After this, a residual diagnostic of model without effects for accident years 4 and 6 is generated in Figure 4.11. This model provides still comparable results with the original model in Figure 4.9. According to Table 4.9, all coefficients are statistically significant except for the last development factor, but it is not reasonable to exclude it from the model due to the precision of prediction.

A figure of predicted, fitted and real claims of the simpler model is not listed because the change compared to Figure 4.10 is hardly noticeable. Nevertheless, the precision of reserve prediction gets worse, from 11172 in original model to 13316 in model without two accident year effects. It is worth mentioning that after excluding only the first mentioned coefficient, prediction was even worse (13642) than after excluding both of them.

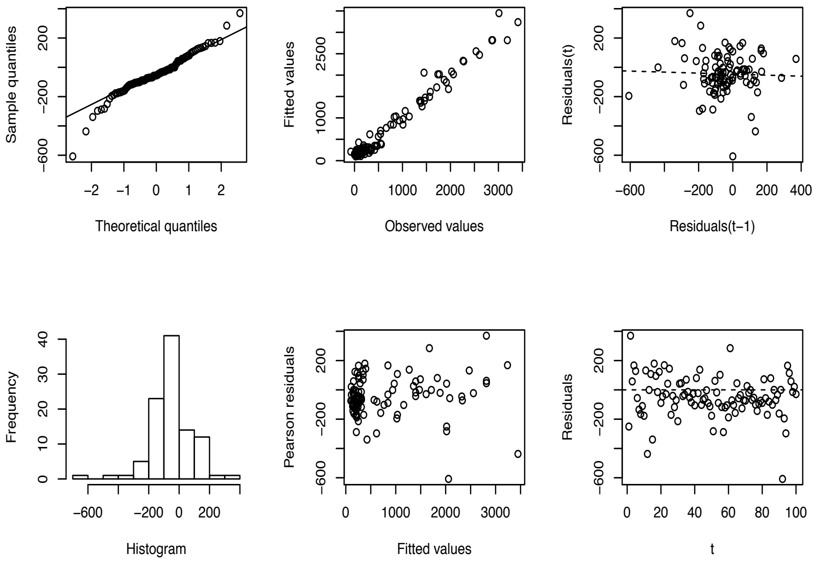

Figure 4.11: Residual diagnostic generated using all rectangle.

Accident year i	Estimation $\widehat{\alpha}_i$ $(\widehat{\gamma})$	P-values Wald test	Development year t	Estimation $\widehat{\beta}_t$	P-values Wald test
1	7.610	$\doteq 0$	2	0.332	$\doteq 0$
2	0.202	$\doteq 0$	3	-0.366	$\doteq 0$
3	0.139	$\doteq 0$	4	-0.873	$\doteq 0$
5	-0.096	$\doteq 0$	5	-1.760	$\doteq 0$
7	-0.187	$\doteq 0$	6	-1.868	$\doteq 0$
8	-0.399	$\doteq 0$	7	-2.257	$\doteq 0$
9	-0.302	$\doteq 0$	8	-2.603	$\doteq 0$
10	-0.314	$\doteq 0$	9	-2.221	$\doteq 0$
			10	-3.841	0.392

Table 4.9: Estimation of coefficient from linear predictor with corresponding P-values.

The purpose of this dataset is to point out, that GLMM is not always the only option, when “variation” between first and second development year occurs. It really depends on the residual diagnostic as well.

On the other hand, it was observed, based on a quite large analysis, that if developments of claims are similar in a certain way through all accident years, i.e., without “variation”, then it would be better to focus a bit more on GEE models. However, only if there is reasonable residual diagnostic.

This type of data can be seen in dataset West Bend Mut Ins. Grp. also from line of business Workers’ compensation. Just as in the previous example, the GEE model with exchangeable correlation structure and variance function equal to one has the best residual diagnostic and predicted reserves are nearest to the real one, from all GEE and GLMM as well as Mack-chain-ladder predictions, see Table 4.10. Only to illustrate how precise our chosen model is, plots of real incremental claims, predicted and fitted values for GEE models with exchangeable correlation structure are shown in Figure 4.12.

Reserves					
Real	Mack	GLMM	Predictions	GEE models	Predictions
45420	42755	Gaussian	42828	AR_Q	42626
		Inv. Gaussian	27418	AR_L	42711
		Gamma	36659	AR_1	42636
				IND_Q	42660
				IND_L	42755
				IND_1	42828
				EX_Q	42660
				EX_L	42957
				EX_1	45043

Table 4.10: Real and predicted reserves.

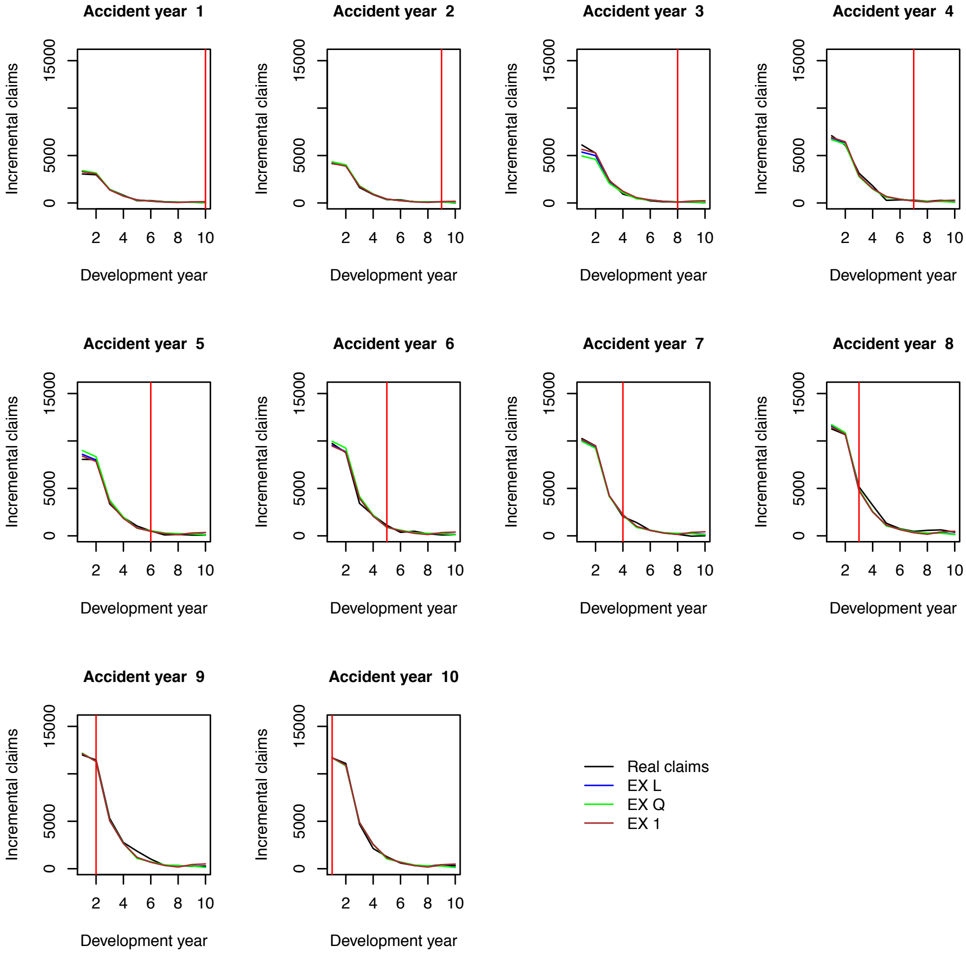

Figure 4.12: Fitted and predicted values vs. real values.

4.4 Problematic example

This section describes a dataset which is not suitable for applying our models. In order to do this, dataset Eveready Ins. Co. from line of business Commercial auto/truck liability/medical is chosen. In Table 4.11, the upper incremental triangle is listed. Already from this table, very volatile behaviour of the data is observed, which is better seen in Figure 4.13.

For example, in accident years 4 and 5, incremental claims have a similar development, but on the other hand, accident year 7 has a totally different pattern. So based on mentioned results from our previous analyses, if we can not decide according to residual diagnostic between GLMM or GEE and “variation” of the data is observed, it may be more appropriate to choose GLMM. Therefore, let’s fit the models in order to generate residual diagnostic.

Accident	Development year j									
year i	1	2	3	4	5	6	7	8	9	10
1	278	469	299	372	376	212	46	6	96	7
2	259	371	257	320	323	182	142	103	1	
3	287	456	368	277	136	278	231	94		
4	218	419	211	419	295	177	259			
5	289	509	262	459	365	149				
6	295	449	198	274	320					
7	303	405	524	690						
8	315	236	296							
9	293	286								
10	456									

Table 4.11: Observed run-off triangle for incremental payments.

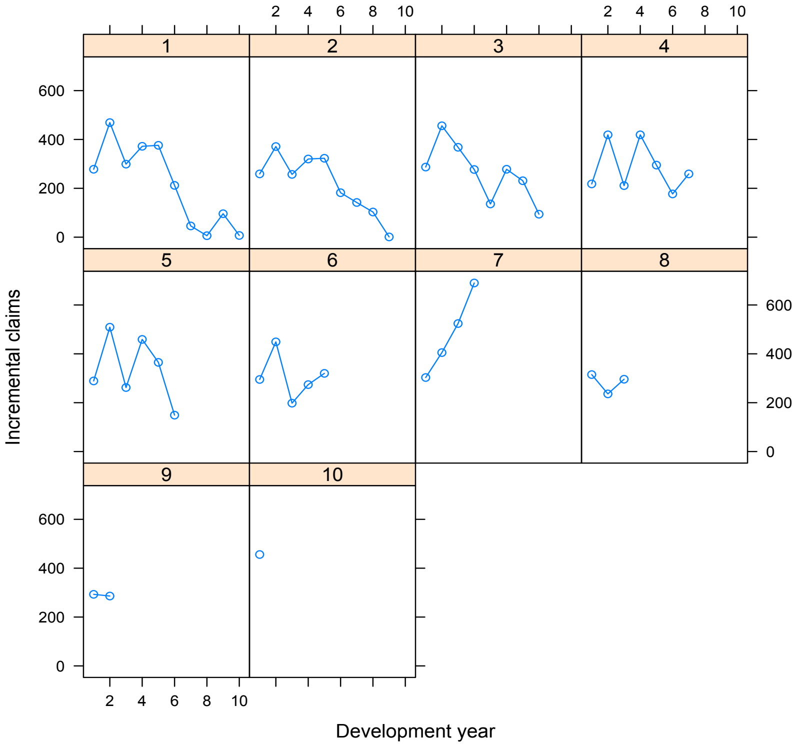

Figure 4.13: Claims development for each accident year.

During the fitting of GLMM to the data, the software did not mention any error in computation. Next, residual diagnostics are generated for all GLMM. However, the obtained results were quite strange, e.g., plot in Figure 4.14, the fitted values with respect to the observed ones, is the same for all three models. Moreover, the fitted values have a range of few numbers, which implies that iterations for all models stopped at the same step and did not converge at the goal. The reason is that the form of the function should be maximized.

In order to fix this problem, different approximations than the Laplace approximation are used, e.g., the adaptive GQ approximation with various number of points per axis for evaluating. None of these approaches make the results better, what implies that GLMM can not be applied to this dataset.

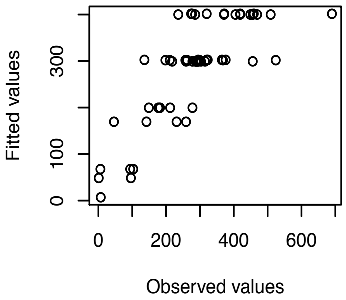

Figure 4.14: Fitted values with respect to observed values.

Furthermore, GEE models are fitted to this data. The whole process of fitting goes well and also some of residual diagnostics seem to be reasonable. According to them, model with $AR(1)$ correlation structure and variance function equal to one is chosen as the most suitable. Residual diagnostic from this model, listed in Figure 4.15, does not immediately imply that this is a useless model. However, a deeper analysis of the range of the residuals and range of the observed values implies that residuals are quite huge. A similar result can be observed from the plot of fitted values with respect to observed ones, where points are not placed around the diagonal much, which means that current model does not fit the data well. This fact is better visible in Figure 4.16, where unlike the previous examples, fitted values are often far away from the real ones.

There is no need to calculate a prediction of reserves because it can be seen mainly from the last accident year in Figure 4.16, that predicted values are far above the real claims. If we use this model, the reserve would be overestimated a lot. Hence, this approach is not very suitable for calculating the prediction of reserve as well.

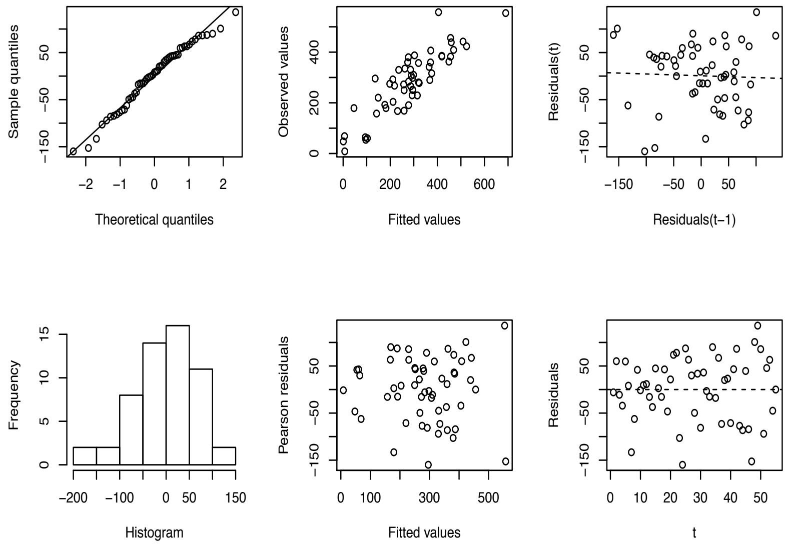

Figure 4.15: Residual diagnostic generated using upper triangle for model *AR*_1.

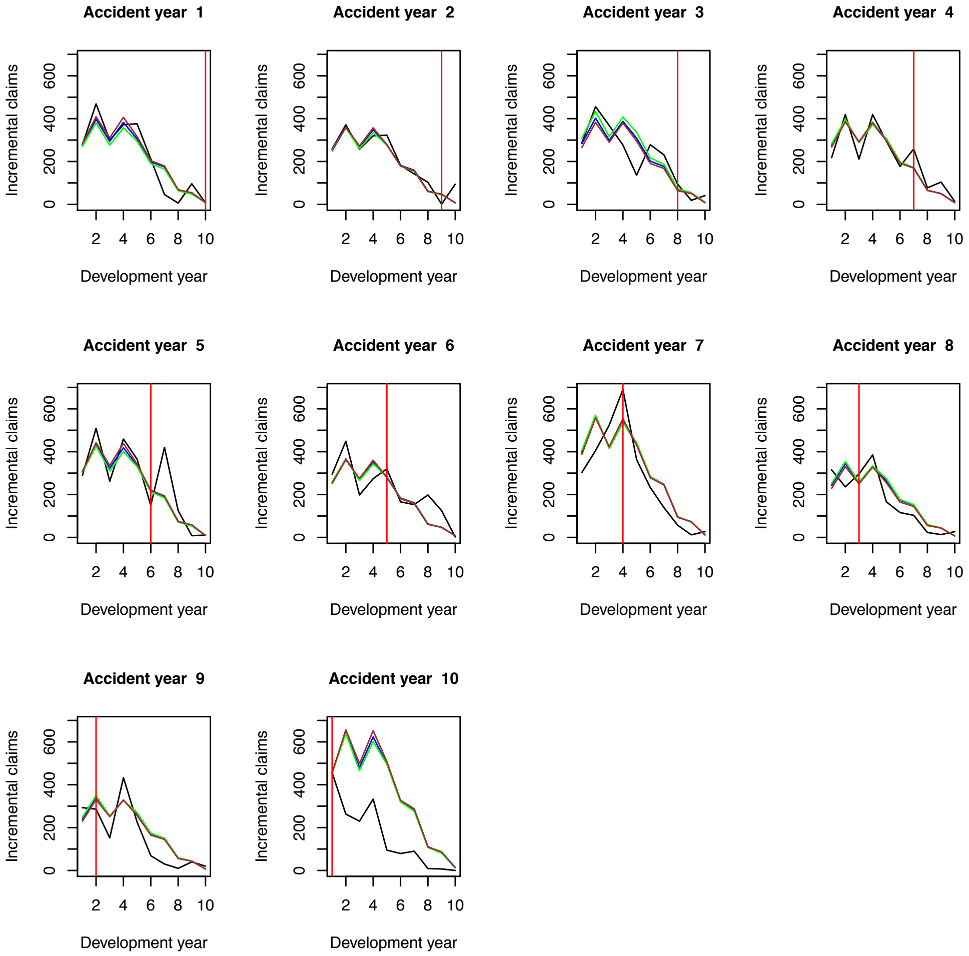

Figure 4.16: Fitted and predicted values vs. real values for models *AR*_1.

4.5 Comparison of results

To sum up, the practical part analyses datasets in order to find out which approach or model is better to use in which case. For the purpose of comparison of the GLMM with GEE models, log-link function is used for both approaches and similar forms of linear predictor are used. The only difference is, that the coefficients for accident years in GEE models are fixed and in GLMM are assumed to be random.

The first part describes the advantages of GLMM applied on volatile data, where Gaussian model is chosen. Subsequently, prediction of reserves using this model was the closest to the real one. In the next part, a dataset with similar data was picked to show the importance of residual diagnostic in model selection. Here, we considered two GEE models but in order to optimize the number of unknown coefficients, GEE model with exchangeable correlation structure and variance function equal to one was chosen. In this case, all GEE models predicted reserves much more accurately than GLMM. Next, the significance of coefficients was tested in order to reduce their number, which has a few important consequences as mentioned in Chapter 3. Using a backward elimination approach and P-values from the Wald test as a criterion, a simpler model without two accident year factors was selected as the most appropriate. This model still has a reasonable diagnostic, however, the prediction of reserve was a bit worse than the original model with all coefficients.

An additional dataset is listed in the GEE section in order to highlight the observation (experience) from the whole database analysis, which favors GEE models when the data is stable in a certain way, as was discussed.

The last but not least part deals with a very volatile dataset, where GLMM failed during the fitting but no errors occurred in the software. This fact was subsequently found out from the residual diagnostic. GEE models were fitted to this dataset as well. However, the chosen model does not fit the data well and the obtained prediction of reserves is way higher than the real one. The purpose of this dataset was to show that it really depends on the behaviour of our data and moreover, our models have their limitations, computational as well as fitting.

Conclusion

The aim of the presented thesis was to implement the GLMM and GEE models, which in contrast to GLM are able to handle with the within-subject correlation, to the problem of claim reserving. We have concentrated on the proper structure of used models and consequently, significant attention has been paid to residual diagnostics and to the suitable choice of the final model in the practical part.

In the first chapter of the thesis, we have presented the whole panel data theory needed for the models definition in general. We have shown the main idea of introducing correlation between the within-subject dependent variables by including additional random variable into the disturbances. The generalization of this idea has led up to the LMM. The advantages of this model, as well as the strengths of GLM, are present in GLMM. Next, we have described the GEE approach, which is able to cope with the within-subject correlation using the working correlation matrix. Our focus has been paid to estimates of unknown regression parameters and unknown working correlation matrix as well as to their properties.

At the beginning of the second chapter, we have introduced the standard notation common in the actuarial science. Furthermore, basic reserving methods have been listed.

The third chapter has joined the first and second chapter in order to synchronize the notation and structure of the models. The goal of this chapter has been to choose the models, which are proper for the insurance data. Attention has been paid to the interpretation and the number of unknown regression parameters that must be estimated. The impact of the inflation and stability of the market has been discussed for this purpose. Due to this, log-link function has been chosen and testing of the coefficients has been made.

Finally, the application of the proposed models on real data has been carried out for the purpose of their analysis and comparison of their performance. All computations have been performed in R software. Firstly, we have made a quite large analysis on the whole database to see in which cases which models are more suitable according to residual diagnostic as well as the precision of fitted values. We have introduced three main datasets. The first one has shown the strength of GLMM, which is the empirical “Bayes estimator” of random intercept. This property works mainly on “volatile” datasets described in GLMM section. The second dataset has had similar variability structure of the data, nevertheless, GEE models has much better residual diagnostic as well as prediction of reserves. This result has implied that firstly, the residual diagnostic should be taken into account. Nevertheless, when we still cannot decide between the GLMM and GEE models, we would recommend GLMM for more “volatile” dataset. The last dataset has shown the limitations: computational for GLMM as well as fitting for GEE models. This fact has confirmed the idea from the quote of George E.P. Box mentioned in the introduction.

Although results from our analysis were satisfactory, the GLMM approach offers much more possibilities in choice of the linear predictor in sense of the random effects. It could be interesting to investigate how these models would fit the data and predict the reserves.

Bibliography

K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58–76, 2007.

B.M. Bolker, M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H.H. Stevens, and J.S.S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135, 2009.

P. J. Diggle, P. Heagerty, K. Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data 2nd ed.* Oxford University Press, Reading, Massachusetts, 2002.

A. J. Dobson. *An introduction to generalized linear models*. CRC Press LLC, Boca Raton, 2002.

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

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

W. H. Greene. *Econometric Analysis*. Prentice Hall, 2002.

J.W. Hardin and J. Hilbe. *Generalized Estimating Equations*. Chapman & Hall, Boca Raton, 2003.

Š. Hudecová and M. Pešta. Modeling dependencies in claims reserving with gee. *Insurance: Mathematics and Economics*, 53:786–794, 2013.

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

K. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

CH. E. McCulloch and S.R. Searle. *Generalized, Linear and Mixed Models*. John Wiley and Sons, Boston, 2001.

G.G. Meyers and P. Shi. Loss reserving data pulled from naic schedule p. `http://exoplanet.eu/catalog.php`, 2011. [Online; posted 28-June-2015].

S. Rabe-Hesketh and A. Skrondal. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 1:1–21, 2002.

C. R. Rao, H. Toutenburg, A. Fieger, C. Heumann, T. Nittner, and S. Scheid. *Linear models: least squares and alternatives 2nd ed.* Wiley Finance, Hoboken, 1999.

S.W. Raudenbush. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *J. Comput. Graph. Statist.*, 9:141–157, 2000.

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

List of Figures

4.1 Claims development for each accident year. 30

4.2 Residual diagnostic generated using upper triangle. 32

4.3 Residual diagnostic generated using all rectangle. 32

4.4 Fitted and predicted values vs. real values. 33

4.5 Pearson residuals vs. fitted values generated using upper triangle. . . . 34

4.6 Claims development for each accident year. 36

4.7 Products of Pearson residuals with respect to their distance within accident year based on the upper triangle. 37

4.8 Residual diagnostic generated using upper triangle for model EX_1. . . 38

4.9 Residual diagnostic generated using all rectangle. 39

4.10 Fitted and predicted values vs. real values. 40

4.11 Residual diagnostic generated using all rectangle. 42

4.12 Fitted and predicted values vs. real values. 44

4.13 Claims development for each accident year. 45

4.14 Fitted values with respect to observed values. 46

4.15 Residual diagnostic generated using upper triangle for model AR_1. . . 47

4.16 Fitted and predicted values vs. real values for models AR_1. 47

List of Tables

2.1 Run-off triangle for incremental claim amounts X_{it}. 23

4.1 Observed run-off triangle for incremental payments. 29
4.2 Lower triangle of incremental payments. 31
4.3 Real and predicted reserves. 34
4.4 Prediction and estimation of coefficient from linear predictor. 35
4.5 Observed run-off triangle for incremental payments. 36
4.6 Lower triangle of incremental payments. 37
4.7 Real and predicted reserves. 41
4.8 Estimation of coefficient from linear predictor with corresponding P-values. 41
4.9 Estimation of coefficient from linear predictor with corresponding P-values. 43
4.10 Real and predicted reserves. 43
4.11 Observed run-off triangle for incremental payments. 45

List of Abbreviations

$\mathbb{R}$	$thesetofrealnumbers$
$\mathbb{N}$	$thesetofnaturalnumbers$
$\mathbf{I}_N$	$diagonalmatrixofoneswithdimensionN$
$\mathbf{1}_T$	$isTdimensionalvectorofones$
$\xrightarrow[n\to\infty]{\mathscr{D}}$	$convergenceindistribution$
$\mathscr{N}$	$normaldistribution$
$\otimes$	$kroneckerproductdefinedasfollows:$
$A \otimes B =$	$\begin{pmatrix} a_{1,1}B & a_{1,2}B & \cdots & a_{1,n}B \\ a_{2,1}B & a_{2,2}B & \cdots & a_{2,n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}B & a_{m,2}B & \cdots & a_{m,n}B \end{pmatrix}$
$A\|B =$	$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_{1,1} & b_{1,2} & \cdots & b_{1,k} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} & b_{2,1} & b_{2,2} & \cdots & b_{2,k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} & b_{m,1} & b_{m,2} & \cdots & b_{m,k} \end{pmatrix}$

A. Source code

In this section, three main R scripts are listed, from which it is easy to construct the whole analysis. The first one is a control script, which goes through the whole database, prepares datasets and calls the functions which form other scripts prepared for each model separately. We describe only two of them because the following scripts are just copies of them with changed distributions, correlation structure or variance function. Hence, the second script constructs the GLMM gamma model with all figures and calculations used in this thesis. The last one is for GEE model with independent correlation structure and variance function equal to one. In this script, all used diagnostics and computations are performed.

A.1 Main control script

```
#pakages
library(lme4)
library(geepack)
library(ChainLadder)
library(tables)

# LoBs: wkcomp_pos, comauto_pos, ppauto_pos
# read data from database
a=read.csv("wkcomp_pos.csv",header=TRUE)
grp.code=unique(a$GRCODE)

ins.line.data=function(g.code){
  b=subset(a,a$GRCODE==g.code)
  name=b$GRNAME
  grpcode=b$GRCODE
  ay=b$AccidentYear
  dev=b$DevelopmentLag
  cum_pdloss=b[,7]
  data.out=data.frame(name,grpcode,ay,dev,cum_pdloss)
  return(data.out)}

# for cycle goes through whole database
#GLMM:   14176  GEE: 715,8559 grp.code (wrk. comp)
#probpematic dataset 11037 (com. auto)
for (f in (grp.code)){

#data into triangles
data=ins.line.data(f)
upper=subset(data,ay+dev<=1998)
triangle=as.triangle(upper,origin="ay", dev="dev",
value="cum_pdloss")
lower=subset(data,ay+dev>1998)
triangle_low=as.triangle(lower,origin="ay", dev="dev",
value="cum_pdloss")
low_inc=cum2incr(triangle_low)

# data from whole rectangle
cely=as.triangle(data,origin="ay", dev="dev", value="cum_pdloss")
cely_inc=cum2incr(cely)
incc=as.vector(t(cely_inc))

# figure of devepment for incremental claims
plot(as.triangle(upper,origin="ay", dev="dev", value="cum_pdloss")
,lattice=TRUE, xlab="Development year",ylab="Incremental claims")

#incremental data upper triangle
pom=upper[5]
for (k in (2:91)){
  if ( is.na(upper[k,3])!='TRUE' & upper[k,3]==upper[k-1,3])
  pom[k,1]=upper[k,5]-upper[k-1,5]
  else if (is.na(upper[k,3])!='TRUE') pom[k,1]=upper[k,5]}
upper[5]=pom

#incremental data from lower triangle
data1=data
pom=data1[5]
pom
for (k in (2:length(data1$cum_pdloss))){
  if ( is.na(data1[k,3])!='TRUE' & data1[k,3]==data1[k-1,3])
  pom[k,1]=data1[k,5]-data1[k-1,5]
  else if (is.na(data1[k,3])!='TRUE') pom[k,1]=data1[k,5]}
data1[5]=pom
lower_inc=subset(data1,ay+dev>1998)

# controll for nonposite incremental values in upper tringle
nonpositive=0
  if ( min(upper[,5])<=0) {
    nonpositive=1}

# input data preparation for model
inc_data=cbind(upper[3],upper[4],upper[5])
colnames(inc_data)[3]="inc_loss"

if (nonpositive == 1 ) {} else {
##############################  GLMMM  ############################

# inicialization of variables
fitall13=rep(0,100)
fitall13=rep(0,100)
fitm12=0
fitm12=0
```

```
rezgaus=0
rezinv=0
rezgam=0

# conditions for datasets which should be skipped with the purpose
# of uninterrupted calculation because they do not converge
if (f == 1538| f== 9466| f == 12297| f==14320 | f==21172) {} else {

# next commands call other R scripts for GLMM models described
# below
source("inverse_Gaussian.R")
m12=inverse11(f)
fitm12=m12$first # just fitted values
fitall12=m12$third # predictions and fitted values
rezinv=m12$fourth}

if (f == 86 | f == 337 | f == 388| f == 1767| f == 2135| f == 2712
| f == 7080 | f==23108 | f==1767) {} else {
source("Gaussian.R")
m13=gausian11(f)
fitm13=m13$first
fitall13=m13$third
rezgaus=m13$fourth}

# an aggregate script plots all models together
source("gamma.R")
m15=gamma11(f)
rezgam=m15$fourth

################################ GEE ##############################
# commands which call scripts for GEE models
source("INDEP_Q.R")
m25=INDEP_Q(f)
fitm5=m25$first
rez1=m25$second

source("INDEP_L.R")
m24=INDEP_L(f)
fitm4=m24$first
rez2=m24$second

source("INDEP_I.R")
m25=INDEP_I(f)
rez3=m25$second
........next GEE models follow

# mack chain ladder
mack <- MackChainLadder(triangle, est.sigma="Mack")
```

```
mackrez=summary(mack)$Total["IBNR",]

# real reserve
realreserve=sum(lower_inc[,5])
realreserve

print("Dataset:")
print(f)
print(c(realreserve,mackrez,rezgaus, rezinv, rezgam))
print(c(rez1, rez2, rez3, rez4, rez5, rez6, rez7, rez8, rez9))  }}
```

A.2 GLMM script

```
# for each model we should have special script like following one
# this script must be saved with name "gamma.R"
gamma11=function(code){
nazov=toString(code)
m1=glmer(inc_loss ~ as.factor(dev) +(1 | ay), data=inc_data,
         family=Gamma("log"), nAGQ=0)
summary(m1)
coef(m1)
fitm1=cbind(inc_data,fitted(m1))
colnames(fitm1)[4]="fit_val"



# fitted values and predictions
c1=c(0,summary(m1)$coef[2:10])
c2=c(exp(coef(m1)[[1]][[1]]))
fit0=c2%o%exp(c1)
fit=as.vector(t(fit0))

# generation of fitted values real claims and predictions only for
#gamma model because it is an aggregation script
pdf(paste("/Users/Michal/Desktop/diplomka/data_analysis/
testovanie_modelov/grafy/",nazov,"GLMM_gamma_prediction_all.pdf"))
par(mfrow=c(3,4))
for (k in (1:10)){
plot(incc[(10*(k-1)+1):(10*k)] ~ inc_data$dev[1:10],type="l",
     , ylab="Incremental claims",xlab= paste("Development year"),
      caption="",xlim=c(1,10),ylim=c(0,max(incc[1:100])))
lines(fit[(10*(k-1)+1):(10*k)]~inc_data$dev[1:10], lwd=1,
col="orange")
lines(fitall12[(10*(k-1)+1):(10*k)]~inc_data$dev[1:10],lwd=1,
col="brown")
lines(fitall13[(10*(k-1)+1):(10*k)]~inc_data$dev[1:10],lwd=1,
col="blue")
  abline(v=(10-k+1),col="red")
title(main=paste("Accident year ", k)
, cex.main=1)
fit }

# legend for plot
plot(0, 0, type = "n", bty = "n", xaxt = "n", yaxt = "n" ,
axes = F, xlab = NA, ylab = NA)
legend("center", c("Real claims", "Gaussian", "Inv. Gaussian",
"Gamma"),
xpd = TRUE, horiz = FALSE, inset = c(0, 0), bty = "n",
 lwd = c(1, 1, 1, 1), col = c("black",  "blue", "brown","orange"),
 cex = 1)
dev.off()

# reserve from GLMM Gamma model
rez=0
for (k in (2:10)){
rez=rez+sum(fit[(10*k-k+2):(10*k)]) }

# residual diagnostic using whole rectangel
pdf(paste("/Users/Michal/Desktop/diplomka/data_analysis/
testovanie_modelov/grafy/",nazov,"GLMM_gamma_residuals.pdf"))
par(mfrow=c(3,3))

res1=(incc-fit)/fit #Pearson residuals
resm1=(incc-fit)
qqnorm(resm1,main="",xlab="Theoretical quantiles",
ylab="Sample quantiles")
qqline(resm1)

plot(fit~incc,xlab="Observed values",ylab="Fitted values")
d1 = length(resm1)
r1=lm(resm1[2:d1]~resm1[1:(d1-1)])
summary(r1)

plot(resm1[2:d1]~resm1[1:(d1-1)],xlab="Residuals(t-1)",
ylab="Residuals(t)")
abline(r1, lty=2)

hist(resm1,main="",xlab="Histogram")
shapiro.test(resm1)

plot(res1 ~ fit, xlab="Fitted values",ylab="Pearson residuals")

plot(resm1, xlab="t",ylab="Residuals")
abline(h=0,lty=2)

dev.off()
return(list(first=fitm1, third=fit, fourth=rez))}
```

A.3 GEE script

```
# this  script is for GEE model INDEP_I and we should make scripts
# like this one for each GEE model
# this script must be saved with name "INDEP_I.R"
INDEP_I=function(code){
nazov=toString(code)

m8=geeglm(inc_loss ~ as.factor(dev) + as.factor(ay), data=inc_data,
family=gaussian("log"), corstr="independence",id=ay)
summary(m8)
coef(m8)
off8=log(inc_data[1,3])
c8=coef(m8)[1]+c(0,coef(m8)[2],coef(m8)[3],coef(m8)[4],coef(m8)[5],
coef(m8)[6],coef(m8)[7], coef(m8)[8],coef(m8)[9],coef(m8)[10])

r8=c(coef(m8)[11],coef(m8)[12],coef(m8)[13],coef(m8)[14],
coef(m8)[15],coef(m8)[16],coef(m8)[17], coef(m8)[18],coef(m8)[19])

#fitted values
fitm7=cbind(exp(c8),exp(r8[1]+c8),exp(r8[2]+c8),exp(r8[3]+c8),
exp(r8[4]+c8),exp(r8[5]+c8),exp(r8[6]+c8),exp(r8[7]+c8),
exp(r8[8]+c8),exp(r8[9]+c8))

# prediction of the ultimate reseves
reserves7=c(fitm7[20],sum(fitm7[29:30]),sum(fitm7[38:40]),
sum(fitm7[47:50]),sum(fitm7[56:60]),sum(fitm7[65:70]),
sum(fitm7[74:80]),sum(fitm7[83:90]),sum(fitm7[92:100]))
reserves7
ultimatereserve7=sum(reserves7)
ultimatereserve7

# this figure is generated only in models when variance function
# is equal to one (like in our case), this is taken as an aggregate
# script and take fitm8 and fitm9 from the control script
pdf(paste("/Users/Michal/Desktop/diplomka/data_analysis/
testovanie_modelov/grafy/",nazov,"GEE_INDEP_I_prediction_all.pdf"))
par(mfrow=c(3,4))
for (k in (1:10)){
plot(incc[(10*(k-1)+1):(10*k)] ~ inc_data$dev[1:10],type="l",
, ylab="incremental claims",xlab= toString(1987+k)
,xlim=c(1,10),ylim=c(0,max(incc[1:100])))
lines(fitm8[(10*(k-1)+1):(10*k)]~inc_data$dev[1:10], lwd=1,
col="blue")
lines(fitm9[(10*(k-1)+1):(10*k)]~inc_data$dev[1:10],lwd=1,
col="green")
lines(fitm7[(10*(k-1)+1):(10*k)]~inc_data$dev[1:10],lwd=1,
col="brown")
```

```
  abline(v=(10-k+1),col="red")}
dev.off()

# residual diagnostic using all rectangle is listed only,
#it is not hard to change it only to upper triangle
pdf(paste("/Users/Michal/Desktop/diplomka/data_analysis/
testovanie_modelov/grafy/",nazov,"GEE_INDEP_I_residuals.pdf"))
par(mfrow=c(3,3))
resm8=(incc-fitm7)
plot(resm8,xlab="",ylab="residuals")
abline(h=0,lty=2)
hist(resm8,main="",xlab="Histogram")
shapiro.test(resm8)
d8 = length(resm8)
r8=lm(resm8[2:d8]~resm8[1:(d8-1)])
qqnorm(resm8,main="")
qqline(resm8)
plot(resm8[2:d8]~resm8[1:(d8-1)],xlab="Residuals(t-1)",
ylab="Residuals(t)")
abline(r8, lty=2)
plot(resm8 ~ fitm7,xlab="fitted values",ylab="residuals")
plot(as.vector(fitm7)~incc,xlab="fitted values",
ylab="observed values")
cor(fitted(m8),inc_data$inc_loss)
plot(as.vector(resm8) ~ rep(c(1:10),10),xlab="accident year",
ylab="residuals")

# this is diagnostic plot for correlation structure only for this
# model
options("scipen"=100, "digits"=3)
mat=matrix(nrow=100, ncol = 10)
pers=vector()
for (j in (1:9)){
for (k in (1:(10-j))){
for (l in ((k+1):(10-j+1)) ){
mat[(k+(9*(j-1))),(l-k)]=(resm8[j,k]*resm8[j,l])}}}
y=c(rep(0,9))
x=c(1:9)
matplot(t(mat),type="p", pch=1, lty=1, lwd=1, col="black",
xlab="|t-k|",ylab=expression(r[it] * r[ik]))
mm=apply(t(mat), 1, function(x) mean(x, na.rm=TRUE))
matpoints(mm,type="p", pch=19, lty=1, lwd=1, col="red")
matlines(c(1:9),c(rep(0,9)), type = "l", lwd = 1,lty=2, pch = NULL)
dev.off()

# in other models, it return data to use them in an aggregate
# script
return(list(first=fitm7,second=ultimatereserve7))}
```