Mixed models for predictive modeling in actuarial science

Katrien Antonio, Yanwei Zhang

Mixed models for predictive modelling in actuarial science

Katrien Antonio ^{*} Yanwei Zhang [†]

September 17, 2012

Outline of this Chapter. We start with a general discussion of mixed (also called multi-level) models and continue with illustrating specific (actuarial) applications of this type of models. Technical details on (linear, generalized, non-linear) mixed models follow: model assumptions, specifications, estimation techniques and methods of inference. Detailed links with useful R packages and SAS procedures are included. Illustrations in this Chapter illustrate the frequentist as well as Bayesian point of view.

1 Mixed models in actuarial science

1.1 What?

Clustered data. **Mixed or multilevel** models are statistical models suitable for the analysis of data structured in nested or non-nested (i.e. cross-classified) **clusters or levels**. These models go beyond the framework of linear and generalized linear models (see Chapter XXX and Chapter XXX), suitable for the analysis of cross-sectional data. In a cross-sectional data set each subject in the sample is observed once. Indicating these objects with i ($i = 1, \dots, m$), this results in a response Y_i and a vector with covariate information $\mathbf{x}_i = (x_{i1} \ x_{i2} \ \dots \ x_{ip})'$. (Generalized) Linear Models [(G)LMs] are directly available for statistical modeling and explain Y_i using the information included in $\mathbf{x}_i$, within an appropriate distributional framework. A cross-sectional analysis assumes independence between subjects i . However, predictive modeling in actuarial science, as in many other statistical disciplines, will confront analysts with data structures going beyond the cross-sectional design dealt with in (G)LMs. Apart from the present chapter, Chapter XXX (on credibility), XXX (on longitudinal data) and XXX (on spatial statistics) in this book include additional examples of clustered data.

^{*}University of Amsterdam and KU Leuven (Belgium), email: k.antonio@uva.nl

[†]TO BE ADDED.

[Here comes a reference to the Chapter on GLMs.]

[Reference to Chapters on credibility, longitudinal data and spatial stats.]

Textbook examples. A standard textbook example of multilevel data is the ‘*students in schools*’ data structure. Extended versions are the ‘*students in classes in schools*’ or ‘*students followed repeatedly over time, in classes in schools*’ examples, where each example is adding an extra level of observations to the data hierarchy. Connecting with the actuarial audience of this book, we consider the example of a collection of vehicles j (with $j = 1, \dots, n_i$) insured under fleets i (with $i = 1, \dots, m$). Let Y_{ij} be the loss random variable observed for vehicle j in fleet i (in a well defined period of exposure). Denote with $x_{1,ij}$ covariate information at vehicle-level (our **level 1**). $x_{1,ij}$ is, for example, the cubic capacity or vehicle age of car j in fleet i . $x_{2,i}$ is a predictor at fleet-level (our **level 2**). $x_{2,i}$ could, for example, refer to the size of the fleet, or the business in which the fleet is operating. The so-called *varying intercepts* model is a basic example of a multilevel model. It combines a linear model at **vehicle-level** (i.e. level 1)

$$Y_{ij} = \beta_i + \beta_{1,0} + x_{1,ij}\beta_{1,1} + \epsilon_{1,ij}, \quad j = 1, \dots, n_i, \quad (1)$$

with a linear model at **fleet-level** (i.e. level 2)

$$\beta_i = \epsilon_{2,i}, \quad i = 1, \dots, m, \quad (2)$$

or, when fleet-specific information is available,

$$\beta_i = x_{2,i}\beta_2 + \epsilon_{2,i}, \quad i = 1, \dots, m. \quad (3)$$

Here $\epsilon_{2,i} \sim (0, \sigma_2^2)^1$ and $\epsilon_{1,ij} \sim (0, \sigma_1^2)$ are mean zero, independent error terms, representing variability (or heterogeneity) at both levels in the data. m is the number of fleets in the study and n_i the number of vehicles registered in fleet i . Written as a **single model equation**, the combination of (1) and, for example, (2), is:

$$Y_{ij} = \beta_{1,0} + \epsilon_{2,i} + x_{1,ij}\beta_{1,1} + \epsilon_{1,ij}. \quad (4)$$

This regression model uses an overall intercept, $\beta_{1,0}$, a fleet-specific intercept, $\epsilon_{2,i}$, a vehicle-level predictor $x_{1,ij}$ with corresponding regression parameter, $\beta_{1,1}$, and an error term $\epsilon_{1,ij}$. We model the fleet-specific intercepts, $\epsilon_{2,i}$, as random variables. This allows to **reflect heterogeneity** between fleets in an efficient way (even for a large number of fleets). Indeed, by assigning a distribution to these error terms, we basically only need an estimate for the unknown parameters in this distribution. Moreover, the random fleet-specific intercepts **structure the dependence** within fleets and allow making fleet-specific predictions, as well as predictions for new fleets (not yet present in the sample). These latter features are - obviously - highly relevant in the context of **predic-**

¹The notation $\epsilon_{2,i} \sim (0, \sigma_2^2)$ implies $E[\epsilon_{2,i}] = 0$ and $\text{Var}[\epsilon_{2,i}] = \sigma_2^2$.

tive modeling. The other regression parameters, $\beta_{1,0}$ and $\beta_{1,1}$, are considered **fixed** (in frequentist terminology); we do not specify a distribution for them. The model in (4) is our first example of a **linear mixed model** ([LMM]). **Mixed** refers to the combination of fixed and random effects, combined in a model specification which is **linear** in the random ($\epsilon_{2,i}$) as well as in the fixed effects ($\beta_{1,0}$ and $\beta_{1,1}$).

Allowing for *varying slopes and intercepts* results in the following model equations

$$Y_{ij} = \beta_{i,0} + x_{1,ij}\beta_{i,1} + \beta_{1,0} + x_{1,ij}\beta_{1,1} + \epsilon_{1,ij}, \quad i = 1, \dots, m, j = 1, \dots, n_i, \quad (5)$$

with

$$\begin{aligned} \beta_{i,0} &= \epsilon_{2,i,0}, \\ \beta_{i,1} &= \epsilon_{2,i,1}. \end{aligned} \quad (6)$$

Written as a single model equation, this second example of a multilevel model becomes

$$Y_{ij} = \beta_{1,0} + \epsilon_{2,i,0} + x_{1,ij}\beta_{1,1} + x_{1,ij}\epsilon_{2,i,1} + \epsilon_{1,ij}. \quad (7)$$

Besides having random intercepts ($\epsilon_{2,i,0}$), the model also allows the effect of predictor $x_{1,ij}$ on the response to vary by fleet. This is modelled here by the random variables $\epsilon_{2,i,1}$.

Main characteristics and motivations. The *varying intercepts* and *varying slopes* examples reveal the essential characteristics of a multilevel model: (1) varying coefficients and (2) a regression model for these varying coefficients (possibly using group-level predictors). Motivations for using multilevel modeling are numerous (see Gelman and Hill (2007)); we illustrate many of them throughout this Chapter. With data often being clustered (e.g. students in schools, students in classes in schools, cars in fleets, panel data, ...), statistical methodology should reflect the structure in the data and use it as relevant information when building statistical models. Using traditional (say (G)LM) regression techniques, the clustering in groups is either ignored (*‘complete pooling’*) or groups are analyzed separately (*‘no pooling’*). With complete pooling differences between clusters are ignored, while the other extreme (*‘no pooling’*) is overfitting the data; even small clusters will get their own regression model. The multilevel model enhances both extremes, e.g. in the varying intercepts model from (4) complete pooling corresponds with $\sigma_2^2 \rightarrow 0$ and $\sigma_2^2 \rightarrow \infty$ with no pooling. Multilevel modeling is a compromise between these two extremes, known as *partial pooling*. In this case, we impose a distributional assumption on $\epsilon_{2,i}$ and estimate σ_2^2 from the data. This allows taking heterogeneity between clusters into account, making appropriate cluster-specific predictions and structuring the dependence between observations belonging to the same cluster. Moreover, predictions related to new clusters become

readily available. Whereas in classical regression cluster-specific indicators can not be included along with cluster-specific predictors, multilevel models allow doing this in a convenient way (see (3)). When specifying regression models at different levels in the data, interactions between explanatory variables at different levels (so-called *cross-level effects*) may appear. The latter is often mentioned as another advantage of multilevel models.

What's in a name?: labels and notation. **Multilevel** models carry many labels in statistical literature. They are sometimes called **hierarchical**, because data are often hierarchically structured (see the students in schools example) and because of the hierarchy in the model specifications. However, non-nested models, with levels structured next to each other, instead of hierarchically nested, can also be analyzed with the multilevel methodology. Multilevel models are also known as **random effects** or **mixed** models, since they combine (a mix of) fixed and random effects. This distinction is only applicable when using frequentist methodology and terminology. A Bayesian analysis treats all regression parameters as random variables, specifying an appropriate prior distribution for each parameter. Besides terminology, mathematical notation can be very different among statistical sources. This should not be a surprise, taking into account that multilevel models can be formulated for basically any number of levels, involving nested and non-nested group effects, predictor information at different levels, and so on. For instance, Gelman and Hill (2007) denote the varying coefficients and varying slopes models in (1)+(3) and (7), respectively, in a more intuitive way:

$$\begin{aligned} Y_i &= \alpha_{j[i]} + \beta x_i + \epsilon_i, & i = 1, \dots, N \\ \alpha_j &= a + bu_j + \eta_j, & j = 1, \dots, m, \end{aligned} \quad (8)$$

and

$$\begin{aligned} Y_i &= \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i, & i = 1, \dots, N \\ \alpha_j &= a_0 + b_0 u_j + \eta_{j1}, & j = 1, \dots, m \\ \beta_j &= \eta_{j2}. \end{aligned} \quad (9)$$

Observations in the data set are indexed with i , where N is the total number of observations. j denotes the fleets in the data set, and $j[i]$ is the fleet to which observation i belongs. x_i refers to covariate information available at vehicle-level (i.e. level 1 in (1)) and u_j refers to covariate information available at fleet-level (i.e. level 2 in (3)).

The notation used from Section 2 on is motivated by generality, and inspired by Frees (2004b). This notation allows writing down model equations in a structured way, with clear reference to the particular level in the data to which the parameter/predictor is attached. Moreover, this notation can be used for any number of levels in a concise

way. Sections 2 and 3 explain the connection between this particular notation and the matrix notation (and corresponding manipulations) often developed in statistical literature on mixed models. When discussing examples, we replace this general notation with a more intuitive one, explicitly referring to the structure of the data under consideration.

1.2 Why?: motivating examples from actuarial science

Research on mixed models originated in bio- and agricultural statistics. For example, the topic of variance components models, a particular example of models with random effects (see Searle et al. (2008)), was studied extensively in the context of animal breeding experiments. The following (non-exhaustive) list of illustrations should convince the reader of the usefulness of mixed models as a modeling tool in actuarial science, with applications ranging from ratemaking to reserving and smoothing.

Illustration 1.1 (Credibility models) *Credibility models – a cornerstone in actuarial mathematics, see Hickman and Heacox (1999) – have a natural and explicit interpretation as special examples of mixed models. Frees et al. (1999) demonstrate this connection, by reinterpreting many credibility models using mixed model parlance. This mapping highly increases the accessibility and usefulness of actuarial credibility models. Indeed, the complete machinery (including computational methods and software) developed for mixed models becomes available for the analysis of actuarial credibility problems. This topic is well developed in Chapter XXX in this book.*

[Reference to
Chapter on
credibility.]

Illustration 1.2 (Workers' Compensation Insurance: frequencies) *The data are from Klugman (1992) (see Scollnik (1996) and Makov et al. (1996) for further discussion). Frequency counts in workers' compensation insurance are observed on a yearly basis for 133 occupation classes followed during 7 years; this is an example of **longitudinal or panel data**. Let Count denote the response variable of interest. Possible explanatory variables are Year and Payroll, a measure of exposure denoting scaled payroll totals adjusted for inflation. Exploratory plots for the raw data (not adjusted for exposure) are in Figure 1. Statistical modeling should take the dependence between observations on the same occupation class into account and reflect the heterogeneity between different classes. In ratemaking (or tarification) an obvious question for this example would be: 'What is the expected number of claims for a risk class in the next observation period, given its observed claims history?'. See Antonio and Beirlant (2007) for further discussion.*

Illustration 1.3 (Workers' Compensation Insurance: losses) *The data set is from the National Council on Compensation Insurance (USA) and contains losses due to permanent partial disability (see Klugman (1992)). 121 occupation or risk classes are observed over a period of 7 years. The variable Loss gives the amount of money paid out*

Figure 1: Boxplots, scatterplot and selection of risk profiles (over time): workers' compensation data (frequencies) from Illustration 1.2.

(on a yearly basis). Possible explanatory variables are *Year* and *Payroll*. Figure 2 shows exploratory plots for the variable *Loss*. The right skewness of the data is apparent. Frees et al. (2001) and Antonio and Beirlant (2007) present mixed models for the pure premium, $PP = \text{Loss} / \text{Payroll}$.

Figure 2: Boxplots, scatterplot and selection of risk profiles (over time): workers' compensation data (losses).

Illustration 1.4 (Hierarchical structures in insurance) With panel data a group of subjects is followed over time, see Illustrations 1.2 and 1.3. This is a basic and widely studied example of hierarchical data. Obviously, more complex structures may occur. Insurance data often come with some kind of **inherent hierarchy**. Motor insurance policies grouped in zip codes within counties within states are one example. Workers' compensation or fire insurance policies operating in similar industries or branches is another one. Consider e.g. the manufacturing versus education branch, with em-

ployees in manufacturing firms indicating larger claims frequencies, and restaurants versus stores, with restaurants having a higher frequency of fire incidents than stores, and so on. A policy holder holding multiple policies (e.g. for theft, motor, flooding, ...), followed over time, within the same company, is an example of a hierarchical data structure studied in the context of **multidimensional credibility** (see Bühlmann and Gisler (2005)). Another detailed multilevel analysis (going beyond the panel data structure) is Antonio et al. (2010). These authors model claim count statistics for vehicles insured under a **fleet** policy. **Fleet policies** are umbrella-type policies issued to customers whose insurance covers more than a single vehicle. The hierarchical or multilevel structure of the data is as follows: vehicles (v) observed over time (t), nested within fleets (f), with policies issued by insurance companies (c). Multilevel models allow for incorporating the hierarchical structure of the data by specifying random effects at vehicle, fleet and company levels. These random effects represent unobservable characteristics at each level. At vehicle level, the missions assigned to a vehicle or unobserved driver behavior may influence the riskiness of a vehicle. At fleet level, guidelines on driving hours, mechanical check-ups, loading instructions and so on, may influence the number of accidents reported. At insurance company level, underwriting and claim settlement practices may affect claims. Moreover, random effects allow a *posteriori* updating of an *a priori* tariff, by taking into account the past performance of vehicle, fleet and company. As such, these models are relevant for a *posteriori* or experience rating with clustered data. See Antonio et al. (2010) and Antonio and Valdez (2012) for further discussion.

Illustration 1.5 (Loss reserving) Zhang et al. (2012) analyze data from the workers' compensation line of business of 10 large insurers, as reported to the National Association of Insurance Commissioners². Common accident years available are from 1988 to 1997. Losses are evaluated at 12-month intervals, with the highest available development age being 120 months. The data have a multilevel structure with losses measured repeatedly over time, among companies and accident years. A plot of the cumulative loss over time for each company clearly shows a nonlinear growth pattern, see Figure 3. Predicting the development of these losses beyond the range of the available data, is the major challenge in loss reserving. Figure 3 reveals that the use of a nonlinear growth curve model is an interesting path to explore. Random effects will be included to structure heterogeneity among companies and between accident years.

Illustration 1.6 (Smoothing) A semiparametric regression model incorporates both parametric as well as nonparametric functional relationships between a response and a set of covariates. In sharp contrast with linear or nonlinear models, nonparametric models do not assume a parametric relationship between covariates and the response *a priori*.

²NAIC is a consortium of state-level insurance regulators in the United States.

Figure 3: Observed growth of cumulative losses for the 10 companies in study.

Rather, such relationships are inferred from the data. As a result, these models are particularly useful when a globally linear pattern is inappropriate or parametric nonlinear curves are difficult to determine. Such nonlinear effect frequently occurs when time related covariates are present, such as driver's age, development lag or years in business of the insured company. For example, in a (G)LM the effect of age of the insured on the number of claims reported is often expressed with a categorical Age covariate. The analyst splits Age in several categories and estimates a regression parameter for each of them. In a nonparametric analysis we model the effect of Age on the response with an unknown, smooth function, in comparison with the piece-wise constant assumption in GLM.

Penalized splines (also called *P-splines*) are popular nonparametric tools that specify the smoothing function as a linear combination of basis functions, in which some coefficients associated with the basis functions are constrained in order to avoid overfitting. That is, they are penalized, or shrunk toward zero, reducing the effective number of coefficients to be estimated. The broad popularity of *P-splines* is largely because they can be written in the form of mixed models with the penalized part of the bases stored in the design matrix $\mathbf{Z}$ (Ruppert et al., 2003; Wood, 2006) so that we can rely on software, diagnostic and inferential tools designed for mixed models directly in fitting *P-splines*, or use a Bayesian implementation of the model to make inference of the full posterior distribution. Of course, hierarchical components can be included in addition to smoothing terms, thus often leading to models that are both intuitively

appealing and structurally flexible when studying practical ratemaking and reserving problems. For example, Figure 4 shows the company-level estimate of the median smoothed incremental losses along with the 50% prediction intervals in a loss triangle analysis based on a four-knot penalized spline on the development lag covariate. More examples of semiparametric models in insurance loss reserving can be found in Antonio and Beirlant (2008) and Zhang and Dukic (2012). Multivariate extensions of penalized splines are available for spatial regression (e.g. in postcode rating). See Chapter XXX for further discussion.

[Reference
to Chapter
on spatial
statistics.]

Figure 4: The plot of the company-level smoother (incremental losses) along with the 50% prediction interval for a loss triangle.

2 Linear mixed models

This Section is based on Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Ruppert et al. (2003), Czado (2004) and Frees (2004b).

2.1 Model assumptions and notation

The basic linear model specifies $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ with $\mathbf{Y}$ an $N \times 1$ vector, $\boldsymbol{\beta}$ a $p \times 1$ vector of regression parameters and $\mathbf{X}$ the $N \times p$ design matrix. N denotes the total number of observations available in the analysis. In traditional statistical parlance, all parameters in $\boldsymbol{\beta}$ are fixed, i.e. no distribution is assigned to them. They are unknown, but fixed constants that should be estimated. In a linear mixed model we start from $\mathbf{X}\boldsymbol{\beta}$, but add $\mathbf{Z}\mathbf{u}$ to it, where $\mathbf{Z}$ is a model matrix, corresponding with a vector of random effects $\mathbf{u}$. A distribution is specified for this random effects vector $\mathbf{u}$ with mean zero and covariance matrix $\mathbf{D}$. As discussed in Section 1 and illustrated below, these random

effects structure between-cluster heterogeneity and within-cluster dependence. All together, textbook notation for linear mixed models is as follows ³

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \\ \mathbf{u} &\sim (\mathbf{0}, \mathbf{D}) \\ \boldsymbol{\epsilon} &\sim (\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{10}$$

with $\boldsymbol{\epsilon}$ a $N \times 1$ vector of error terms with covariance matrix $\boldsymbol{\Sigma}$ (see below for examples), which is independent of $\mathbf{u}$. This is the *hierarchical* specification of a linear mixed model. For given $\mathbf{u}$ the conditional mean and variance are

$$\begin{aligned} E[\mathbf{Y}|\mathbf{u}] &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \\ \text{Var}[\mathbf{Y}|\mathbf{u}] &= \boldsymbol{\Sigma}. \end{aligned} \tag{11}$$

The combined, unconditional or *marginal* model states

$$\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V} := \mathbf{Z}\mathbf{D}\mathbf{Z}' + \boldsymbol{\Sigma}), \tag{12}$$

showing that fixed effects enter the (implied) mean of $\mathbf{Y}$ and random effects structure the (implied) covariance matrix of $\mathbf{Y}$.

Usually, normality is assumed for $\mathbf{u}$ and $\boldsymbol{\epsilon}$, thus

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{pmatrix} \right). \tag{13}$$

With these distributional assumptions the hierarchical LMM becomes

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \boldsymbol{\Sigma}) \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{D}). \end{aligned} \tag{14}$$

This implies the marginal model $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, but not vice versa. When interest is only in the fixed effects parameters $\boldsymbol{\beta}$ the marginal model can be used. With explicit interest in $\boldsymbol{\beta}$ and $\mathbf{u}$ the specification in (10) and (14) should be used.

Illustrations 2.1 and 2.2 below focus on particular examples of 2 and 3 level data and explain in detail the structure of vectors and matrices in (10) and (12).

Illustration 2.1 (A 2-level model for longitudinal data.) Y_{ij} represents the j th measurement on a subject i (with $i = 1, \dots, m$ and $j = 1, \dots, n_i$). m is the number of subjects under consideration and n_i the number of observations registered on subject i . x_{ij}

³The notation $\mathbf{u} \sim (\mathbf{0}, \mathbf{D})$ implies $E[\mathbf{u}] = \mathbf{0}$ and $\text{Var}[\mathbf{u}] = \mathbf{D}$.

$(p \times 1)$ is a column vector with fixed effects' covariate information from observation j on subject i . Correspondingly, z_{ij} ($q \times 1$) is a column vector with covariate information corresponding with random effects. β ($p \times 1$) is a column vector with fixed effects parameters and u_i ($q \times 1$) is a column vector with random effects regression parameters. These are subject-specific and allow to model heterogeneity between subjects. The combined model is

$$Y_{ij} = \underbrace{x'_{ij}\beta}_{\text{fixed}} + \underbrace{z'_{ij}u_i}_{\text{random}} + \underbrace{\epsilon_{ij}}_{\text{random}}. \quad (15)$$

The distributional assumptions for the random parts in (15) are

$$\begin{aligned} u_i &\sim (0, G) & G &\in \mathbb{R}^{q \times q} \\ \epsilon_i &\sim (0, \Sigma_i) & \Sigma_i &\in \mathbb{R}^{n_i \times n_i}. \end{aligned} \quad (16)$$

The covariance matrix G is left unspecified, i.e. no particular structure is implied. Various structures are available for Σ_i . Very often just a simple diagonal matrix is used: $\Sigma_i := \sigma^2 I_{n_i}$. However, when the inclusion of random effects is not enough to capture the dependence between measurements on the same subject, we can add serial correlation to the model and specify Σ_i as non-diagonal (e.g. unstructured, Toeplitz or autoregressive structure, see Verbeke and Molenberghs (2000) for more discussion). $u_1, \dots, u_m, \epsilon_1, \dots, \epsilon_m$ are independent. Typically, normality is assumed for both vectors, as in (14). In vector notation we specify

$$\begin{aligned} Y_i &= X_i \beta + Z_i u_i + \epsilon_i, & i &= 1, \dots, m, \\ u_i &\sim (0, G) \\ \epsilon_i &\sim (0, \Sigma_i), \end{aligned} \quad (17)$$

where

$$X_i := \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times p}, \quad Z_i = \begin{pmatrix} z'_{i1} \\ \vdots \\ z'_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times q}, \quad Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times 1}. \quad (18)$$

Combining all subjects or clusters $i = 1, \dots, m$, (10) is the matrix formulation of this

LMM for longitudinal data (with $N = \sum_{i=1}^m n_i$ the total number of observations)

$$\begin{aligned} \mathbf{Y} &= \begin{pmatrix} \mathbf{Y}_1 \\ \dots \\ \mathbf{Y}_m \end{pmatrix} \in \mathbb{R}^{N \times 1}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \in \mathbb{R}^{N \times p}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix} \in \mathbb{R}^{N \times 1}, \\ \mathbf{Z} &= \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0}_{n_1 \times q} & \dots & \mathbf{0}_{n_1 \times q} \\ \mathbf{0}_{n_2 \times q} & \mathbf{Z}_2 & & \\ \vdots & & \ddots & \\ \mathbf{0}_{n_m \times q} & & & \mathbf{Z}_m \end{pmatrix} \in \mathbb{R}^{N \times (m \cdot q)}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{pmatrix} \in \mathbb{R}^{(m \cdot q) \times 1}. \end{aligned} \quad (19)$$

The covariance matrix of the combined random effects vector $\mathbf{u}$ on the one hand, and the combined residual vector $\boldsymbol{\epsilon}$ on the other hand, are specified as:

$$\mathbf{D} = \begin{pmatrix} \mathbf{G} & & \\ & \ddots & \\ & & \mathbf{G} \end{pmatrix} \in \mathbb{R}^{m \cdot q \times m \cdot q}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_m \end{pmatrix} \in \mathbb{R}^{N \times N}. \quad (20)$$

Covariance matrix $\mathbf{V}$ in this particular example is block diagonal and given by

$$\begin{aligned} \mathbf{V} &= \mathbf{Z} \mathbf{D} \mathbf{Z}' + \boldsymbol{\Sigma} \\ &= \begin{pmatrix} \mathbf{Z}_1 \mathbf{G} \mathbf{Z}_1' + \boldsymbol{\Sigma}_1 & \dots & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Z}_m \mathbf{G} \mathbf{Z}_m' + \boldsymbol{\Sigma}_m \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{V}_1 & & \\ & \ddots & \\ & & \mathbf{V}_m \end{pmatrix}, \end{aligned} \quad (21)$$

with $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \boldsymbol{\Sigma}_i$.

Illustration 2.2 (A 3-level example.) Y_{ijk} is the response variable of interest, as observed for, say, vehicle k , insured in fleet j by insurance company i . At vehicle level (or: level 1) we model this response as:

$$Y_{ijk} = z'_{1,ijk} \boldsymbol{\beta}_{ij} + \mathbf{x}'_{1,ijk} \boldsymbol{\beta}_1 + \epsilon_{1,ijk}. \quad (22)$$

Hereby, predictors $z_{1,ijk}$ and $x_{1,ijk}$ may depend on insurance company, fleet or vehicle. $\boldsymbol{\beta}_1$ is a vector of regression parameters which will not vary by company nor fleet; they are fixed effects regression parameters. Parameters $\boldsymbol{\beta}_{ij}$ vary by company and fleet. We

model them in a level 2-equation:

$$\beta_{ij} = \mathbf{Z}_{2,ij}\gamma_i + \mathbf{X}_{2,ij}\beta_2 + \epsilon_{2,ij}. \quad (23)$$

$\mathbf{X}_{2,ij}$ and $\mathbf{Z}_{2,ij}$ may depend on company or fleet, but not on the insured vehicle. The regression parameters in γ_i are company-specific and modeled in (24):

$$\gamma_i = \mathbf{X}_{3i}\beta_3 + \epsilon_{3i}, \quad (24)$$

where the predictors in $\mathbf{X}_{3i}$ may depend on company, but not on fleet or vehicle. The combined level 1, 2 and 3 models lead to the following model specification:

$$\begin{aligned} Y_{ijk} &= \mathbf{z}'_{1,ijk}(\mathbf{Z}_{2,ij}(\mathbf{X}_{3i}\beta_3 + \epsilon_{3i}) + \mathbf{X}_{2,ij}\beta_2 + \epsilon_{2,ij}) + \mathbf{x}'_{1,ijk}\beta_1 + \epsilon_{1,ijk} \\ &= \mathbf{x}'_{ijk}\beta + \mathbf{z}'_{ijk}\mathbf{u}_{ij} + \epsilon_{1,ijk}, \end{aligned} \quad (25)$$

where $\mathbf{x}'_{ijk} = (\mathbf{x}'_{1,ijk} \ \mathbf{z}'_{1,ijk}\mathbf{X}_{2,ij} \ \mathbf{z}'_{1,ijk}\mathbf{Z}_{2,ij}\mathbf{X}_{3i})$, $\beta = (\beta'_1 \ \beta'_2 \ \beta'_3)'$, $\mathbf{z}'_{i,j,k} = (\mathbf{z}'_{1,i,j,k} \ \mathbf{z}'_{1,i,j,k}\mathbf{Z}_{2,ij})$ and $\mathbf{u}_{ij} = (\epsilon'_{2,ij} \ \epsilon'_{3,i})'$. Formulating this 3-level model in matrix notation follows from stacking all observations Y_{ijk} .

More examples of LMM specifications are in McCulloch and Searle (2001). A standard notation for a k -level model is in Frees (2004b) (Appendix 5A).

2.2 The structure of random effects

Since the random effects $\mathbf{u}$ are often corresponding to factor predictors, the design matrix $\mathbf{Z}$ is often highly sparse, with a high proportion of elements to be exactly zero. Moreover, the covariance matrix $\mathbf{D}$ is highly structured and depends on some parameter θ that is to be estimated. For example, Figure 5 shows the design matrices for four commonly encountered scenarios ⁴

- **Single random effect per level.** This is the simplest yet most common case where u_j corresponds to the j_{th} level of a single grouping factor. For example, we may have the state indicator in the model and each state has its own intercept. Figure 5 (a) shows the transposed design matrix $\mathbf{Z}'$ in this case, for a factor with 12 levels. A distinctive feature is that each column in the figure, that is, each row of the design matrix $\mathbf{Z}$ has exactly one element that is not zero. Thus, there are exactly n non-zero values in the design matrix. Illustration 3.3, model (89), is an example of this kind of structure.

⁴The programming code in XXX demonstrates how the design matrix, $\mathbf{Z}$, of a specific mixed model can be pictured.

Figure 5: Visualization of the transposed design matrix of the random effects for (a) single random effect per level; (b) multiple random effects per level; (c) nested random effects and (d) crossed random effects. Only colored cells are filled with non-zero values.

- **Multiple random effects per level.** Another common case is that the model has both random intercepts and random slopes that vary by some grouping factor. For example, each state in the model has its own intercept and slope. Figure 5 (b) shows the associated design matrix. We see that each column of Z' has two non-zeros that equal-distanced across the columns. In general, the multiple random effects are correlated, and so the random component D is not diagonal. Illustration 3.3, model (90), is an example of a mixed model with both random intercepts and random slopes.
- **Nested random effects.** In the nested classification, some levels of one factor occur only within certain levels of a first factor. For example, we may have observations within each county, and then the counties within each state. The county from state A never occurs for state B, so counties are nested within states, forming a hierarchical structure. Figure 5 (c) shows such a situation. Antonio et al. (2010) is an example of this type of structuring.
- **Crossed random effects.** This happens when each level of each factor may occur with each level of each other factor. For example, we may have both state and car make in the model, cars of different makes can occur with each state. Figure 5 (d) shows such a situation where two crossed factors are included, one with 5 levels and the other with 12 levels. The credit insurance example in Antonio and Beirlant (2007) is an example of crossed random effects.

2.3 Parameter estimation, inference and prediction

Mixed models use a combination of fixed effects regression parameters, random effects and covariance matrix parameters (also called: *variance components*). For example, in the varying intercepts example from (1) and (2), $\beta_{1,0}$ and $\beta_{1,1}$ are regression parameters corresponding with fixed effects, σ_1^2 and σ_2^2 are variance components and $\epsilon_{2,i}$ ($i = 1, \dots, m$) are the random effects. We will use standard statistical methodology, like maximum likelihood, to *estimate parameters* in a LMM. For the random effects we apply statistical knowledge concerning *prediction* problems, see McCulloch and Searle (2001) (Chapter 9) for an overview. The difference in terminology stems from the non-randomness of the parameters versus the randomness of the random effects.

We first derive an estimator for the fixed effects parameters in β (and a predictor for the random effects in u), under the assumption of known covariance parameters in V (see (12)).

Estimating β . The Generalized Least Squares ([GLS]) estimator – which coincides with the maximum likelihood estimator ([MLE]) under normality (as in (14)) – of β is:

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y. \quad (26)$$

See Frees (2004b) or Czado (2004) for a formal derivation of this result.

Predicting u . In the sense of minimal Mean Squared Error of Prediction ([MSEP]) the best predictor ([BP]) of u is the conditional mean $E[u|Y]$. This predictor obviously requires knowledge of the conditional distribution $u|Y$. The BP is often simplified by restricting the predictor to be a linear function of Y : the Best Linear Predictor ([BLP]). The BLP of a random vector u is

$$\text{BLP}[u] = \hat{u} = E[u] + CV^{-1}(Y - E[Y]), \quad (27)$$

where $V = \text{Var}(Y)$ and $C = \text{Cov}(u, Y')$. $\text{BP}(u)$ and $\text{BLP}(u)$ are unbiased, in the sense that their expected value equals $E[u]$. Normality is not required in BP or BLP, but with $(Y \ u)$ multivariate normally distributed, the BP and BLP coincide. See McCulloch and Searle (2001) (Chapter 9) for more details.

In the context of the LMM sketched in (14) the predictor of u is usually called the Best Linear Unbiased Predictor ([BLUP]). Robinson (1991) describes several ways to

derive this BLUP. For instance, under normality assumptions:

$$\begin{aligned} \text{Cov}(\mathbf{Y}, \mathbf{u}') &= \text{Cov}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \mathbf{u}') \\ &= \text{Cov}(\mathbf{X}\boldsymbol{\beta}, \mathbf{u}') + \mathbf{Z}\text{Var}(\mathbf{u}, \mathbf{u}') + \text{Cov}(\boldsymbol{\epsilon}, \mathbf{u}') \\ &= \mathbf{Z}\mathbf{D}, \end{aligned}$$

which leads to the multivariate normal distribution

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{u} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{Z}\mathbf{D} \\ \mathbf{D}\mathbf{Z}' & \mathbf{D} \end{pmatrix} \right). \quad (28)$$

Using either properties of this distribution⁵ or the result in (27) the BLUP of $\mathbf{u}$ follows:

$$\text{BLUP}(\mathbf{u}) := \hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \quad (29)$$

Of course, (29) relies on the (unknown) vector of fixed effects $\boldsymbol{\beta}$, as well as on unknown covariance parameters in $\mathbf{V}$. Replacing both with their estimates, we call the BLUP an empirical or estimated BLUP. Estimated BLUPs are confronted with multiple sources of variability: variability from the estimation of $(\boldsymbol{\beta}, \mathbf{u})$ and from the estimation of $\mathbf{V}$. Histograms and scatter plots of components of $\hat{\mathbf{u}}$ are often used to detect outlying clusters, or to visualize between-cluster heterogeneity.

A unified approach: Henderson's justification. Maximizing the joint log likelihood of $(\mathbf{Y}', \mathbf{u}')'$ (see assumptions (14)) with respect to $(\boldsymbol{\beta}, \mathbf{u})$ leads to Henderson's mixed model equations:

$$\begin{aligned} f(\mathbf{y}, \mathbf{u}) &= f(\mathbf{y}|\mathbf{u}) \cdot f(\mathbf{u}) \\ &\propto \exp \left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \right) \cdot \exp \left(-\frac{1}{2}\mathbf{u}' \mathbf{D}^{-1}\mathbf{u} \right) \end{aligned} \quad (30)$$

It is therefore enough to minimize

$$Q(\boldsymbol{\beta}, \mathbf{u}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}' \mathbf{D}\mathbf{u}, \quad (31)$$

⁵Namely: with $\mathbf{X} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim N \left(\begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_Z \end{pmatrix} \right)$ we know $\mathbf{Z}|\mathbf{Y} \sim N(\boldsymbol{\mu}_{Z|Y}, \boldsymbol{\Sigma}_{Z|Y})$ where $\boldsymbol{\mu}_{Z|Y} = \boldsymbol{\mu}_Z + \boldsymbol{\Sigma}_{ZY}\boldsymbol{\Sigma}_Y^{-1}(\mathbf{Y} - \boldsymbol{\mu}_Y)$ and $\boldsymbol{\Sigma}_{Z|Y} = \boldsymbol{\Sigma}_Z - \boldsymbol{\Sigma}_{ZY}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{YZ}$.

which corresponds to solving the set of equations

$$\begin{aligned} \frac{\partial}{\partial \beta} Q(\beta, u) &= 0 \text{ and } \frac{\partial}{\partial u} Q(\beta, u) = 0 \\ \Leftrightarrow \begin{pmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{X} & \mathbf{X}'\Sigma^{-1}\mathbf{Z} \\ \mathbf{Z}'\Sigma^{-1}\mathbf{X} & \mathbf{Z}'\Sigma^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ \tilde{u} \end{pmatrix} &= \begin{pmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{y} \\ \mathbf{Z}'\Sigma^{-1}\mathbf{y} \end{pmatrix}. \end{aligned} \quad (32)$$

(26) and (29) solve this system of equations.

More on prediction. With $\hat{\beta}$ from (26) and $\hat{u}$ from (29), the profile of cluster i is predicted by

$$\begin{aligned} \hat{\mathbf{Y}}_i &:= \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{u}_i \\ &= \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\beta}) \\ &= \Sigma_i\mathbf{V}_i^{-1}\mathbf{X}_i\hat{\beta} + (\mathbf{I}_{n_i} - \Sigma_i\mathbf{V}_i^{-1})\mathbf{Y}_i, \end{aligned} \quad (33)$$

using $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \Sigma_i$ and n_i the cluster size. $\hat{\mathbf{Y}}_i$ is a weighted mean of the global profile $\mathbf{X}_i\hat{\beta}$ and the data observed on cluster i , $\mathbf{Y}_i$. $\hat{\mathbf{Y}}_i$ is a so-called *shrinkage estimator*. Actuaries will recognize a credibility formula in (33).

The prediction of a future observation is discussed in detail in Frees (2004a) (Section 4.4). The case of non-diagonal residual covariance matrices Σ_i requires special attention. For instance, with panel data the BLUP for Y_{i,T_i+1} is $\mathbf{x}_{i,T_i+1}'\beta + \mathbf{z}_{i,T_i+1}'\hat{u}_i + \text{BLUP}(\epsilon_{i,T_i+1})$. From (27) we understand that the last term in this expression is zero when $\text{Cov}(\epsilon_{i,T_i+1}, \epsilon_i) = \mathbf{0}$. This is not the case when serial correlation is taken into account. Chapter XXX of this book (on *Credibility and Regression Modeling*) carefully explains this kind of prediction problems.

[Here we connect with Chapter 17 from the book.]

The parameters or variance components used in $\mathbf{V}$ are in general unknown and should be estimated from the data. With θ the vector of unknown parameters used in $\mathbf{V} = \mathbf{Z}\mathbf{D}(\theta)\mathbf{Z}' + \mathbf{D}(\theta)$, the log-likelihood for (β, θ) is (with c a constant)

$$\begin{aligned} \ell(\beta, \theta) &= \log \{L(\beta, \theta)\} \\ &= -\frac{1}{2} \left(\ln |\mathbf{V}(\theta)| + (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}(\theta)^{-1} (\mathbf{y} - \mathbf{X}\beta) \right) + c. \end{aligned} \quad (34)$$

Maximizing (34) with respect to β and with θ fixed, we get

$$\hat{\beta}(\theta) = (\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{Y}. \quad (35)$$

We obtain the so-called *profile log-likelihood* by replacing β in (34) with $\hat{\beta}$ from (35)

$$\begin{aligned}\ell_p(\theta) &:= \ell(\hat{\beta}, \theta) \\ &= -\frac{1}{2} \left\{ \ln |\mathbf{V}(\theta)| + (\mathbf{y} - \mathbf{X}\hat{\beta}(\theta))' \mathbf{V}(\theta)^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}(\theta)) \right\}.\end{aligned}\quad (36)$$

Maximizing this profile log-likelihood with respect to θ gives the maximum likelihood estimates $\hat{\theta}_{\text{MLE}}$ of the variance components in θ .

With LMMs Restricted (or Residual) maximum likelihood (REML) is a popular alternative to estimate θ . REML accounts for the degrees of freedom used for fixed effects estimation. McCulloch and Searle (2001) (Section 6.10) is an overview of important arguments in the discussion '*ML versus REML?*'. For example, estimates with REML (for balanced data) are minimal variance unbiased under normality⁶, and are invariant to the value of β . The REML estimation of θ is based on the marginal log-likelihood obtained by integrating out the fixed effects in β :

$$\ell_r(\theta) := \ln \left(\int L(\beta, \theta) d\beta \right), \quad (38)$$

where (see Czado (2004))

$$\begin{aligned}\int L(\beta, \theta) d\beta &= \int \frac{1}{(2\pi)^{N/2}} |\mathbf{V}(\theta)|^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}(\theta)^{-1} (\mathbf{y} - \mathbf{X}\beta) \right) d\beta \\ &\vdots \\ &= \ell_p(\theta) - \frac{1}{2} \ln |\mathbf{X}' \mathbf{V}(\theta)^{-1} \mathbf{X}| + \text{constants}.\end{aligned}\quad (39)$$

2.3.1 Standard errors and inference

Estimation of standard errors. In the marginal model $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V}(\theta))$, the covariance of $\hat{\beta}$ in (26) is

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}' \mathbf{V}^{-1}(\theta) \mathbf{X})^{-1}, \quad (40)$$

where $\text{Cov}(\mathbf{Y}) = \mathbf{V}(\theta)$ is used. Replacing the unknown θ with its ML or REML estimate $\hat{\theta}$ and using $\hat{\mathbf{V}} := \mathbf{V}(\hat{\theta})$, a natural estimate for $\text{Cov}(\hat{\beta})$ is $(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}$. However, this estimate ignores the extra variability originating from the estimation of θ .

⁶A well known example of '*REML versus ML*' considers the case of a random sample $X_1, \dots, X_N \sim N(\mu, \sigma^2)$. The resulting estimators for the unknown variance σ^2 are

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2, \quad \hat{\sigma}_{\text{REML}}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2, \quad (37)$$

with $\bar{X}$ the sample mean. The REML estimator is unbiased for σ^2 . The $(N-1)$ in $\hat{\sigma}_{\text{REML}}^2$ accounts for the estimation of μ via $\bar{X}$.

Kacker and Harville (1984) (among others) discuss attempts to quantify this extra variability through approximation, but only a fully Bayesian analysis allows to account for all sources of variability (see Section 5).

The covariance of the empirical BLUP in (29) is equal to

$$\begin{aligned}\text{Cov}(\hat{\mathbf{u}}) &= \text{Cov}(\mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})) \\ &= \mathbf{DZ}' \left\{ \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X} \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{V}^{-1} \right\} \mathbf{ZD}.\end{aligned}\quad (41)$$

However, the estimator in (41) ignores the variability in the random vector $\mathbf{u}$. Therefore, as suggested by Laird and Ware (1982), inference for $\mathbf{u}$ is usually based on $\text{Cov}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{D} - \text{Cov}(\hat{\mathbf{u}})$. Estimates of the precision of other predictors involving $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are based on

$$\text{Cov} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix}, \quad (42)$$

and are available in McCulloch and Searle (2001) (Section 9.4 (c)). Accounting for the variability induced by estimating the variance components $\boldsymbol{\theta}$ would require – once again – a fully Bayesian analysis. Using Bayesian statistics posterior credible intervals of cluster-specific effects follow immediately. These are useful to understand the between-cluster heterogeneity present in the data.

With respect to inference for the variance components in $\boldsymbol{\theta}$, McCulloch and Searle (2001) (Section 6.12) derive the information matrix in the general model using the normality assumptions as in (14).

Inference. We consider testing a set of s ($s \leq p$) hypotheses concerning the fixed effects parameters in $\boldsymbol{\beta}$

$$\begin{aligned}H_0 : \mathbf{C}\boldsymbol{\beta} &= \boldsymbol{\zeta} \\ \text{versus } H_1 : \mathbf{C}\boldsymbol{\beta} &\neq \boldsymbol{\zeta}.\end{aligned}\quad (43)$$

The Wald test statistic

$$[\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\zeta}]' [\mathbf{C}\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{C}'] [\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\zeta}] \quad (44)$$

is approximately χ_s^2 distributed. With $\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$ the log-likelihood obtained with ML in the restricted model (i.e. under H_0) and $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ the log-likelihood with ML in the unrestricted model, the likelihood ratio test statistic ([LRT]) for nested models

$$-2[\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}) - \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})], \quad (45)$$

is approximately χ_s^2 distributed. Estimation should be done with ML instead of REML, since REML maximizes the likelihood of linear combinations of $\mathbf{Y}$ that do not depend on β .

Testing the necessity of random effects requires a hypothesis test involving the variance components. For example, in the varying intercepts model from (4), we want to investigate whether the intercepts of different subjects are significantly different. This corresponds with

$$H_0 : \sigma_2^2 = 0 \quad \text{versus} \quad H_1 : \sigma_2^2 > 0. \quad (46)$$

However, because 0 is on the boundary of the allowed parameter space for σ_2^2 , the likelihood ratio test statistic should not be compared with a χ_1^2 distribution, but with a mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. When testing a hypothesis involving s fixed effects parameters and one variance component, the reference distribution is $\frac{1}{2}\chi_s^2 + \frac{1}{2}\chi_{s+1}^2$. When more variance components are involved, the complexity of this problem highly increases, see Ruppert et al. (2003) and related work from these authors.

3 Generalized linear mixed models

Generalized Linear Models ([GLMs]) have numerous applications in actuarial science, ranging from ratemaking over loss reserving to mortality modeling. See Haberman and Renshaw (1996) for an overview. Chapter XXX of this book explains in detail the use of GLMs with cross-sectional data. In what follows we present an extension of these models towards clustered or hierarchical data.

[Reference to the chapter on GLMs.]

3.1 A recap of GLMs

A GLM is a regression model specified within the distributional framework of the exponential family. A member of this family has a density of the following form

$$f_Y(y) = \exp \left(\frac{y\theta - \psi(\theta)}{\phi} + c(y, \phi) \right). \quad (47)$$

$\psi(\cdot)$ and $c(\cdot)$ are known functions, θ is the natural and ϕ the scale parameter. Using vector notation the following relations hold

$$\mu = E[\mathbf{Y}] = \psi'(\theta) \quad \text{and} \quad \text{Var}[\mathbf{Y}] = \phi\psi''(\theta) = \phi V(\mu), \quad (48)$$

where derivatives are with respect to θ and $V(\cdot)$ is the variance function. The latter function captures the relationship, if any exists, between the mean and variance of Y .

GLMs provide a way around transforming data, by specifying a transformation of the mean as a linear function of some regression parameters

$$g(\mu) = \eta = \mathbf{X}\beta, \quad (49)$$

with $\beta = (\beta_1, \dots, \beta_p)'$ the vector of regression parameters and $\mathbf{X}$ ($N \times p$) the design matrix. g is the link function and η the so-called linear predictor. Estimates for β follow by solving the maximum likelihood equations with an iterative numerical technique (such as Newton-Raphson). Likelihood ratio and Wald tests are available for inference purposes. If the scale parameter ϕ is unknown, it can be estimated either by maximum likelihood or by dividing the deviance or Pearson's chi-square statistic by its degrees of freedom.

3.2 Model families for hierarchical non-Gaussian data

In Section 2 we explained the connection between the marginal and hierarchical interpretation of a LMM. This feature is a consequence of the nice properties of the multivariate normal distribution, but it will no longer exist when outcomes are of non-Gaussian type. This implies - among others - that the fixed effects β have different interpretation in so-called marginal versus random effects models for clustered non-normal data. Estimates obtained with one of both model families may differ substantively. Molenberghs and Verbeke (2005) distinguish 3 model families for handling non-Gaussian clustered data: *marginal*, *conditional* and *subject-specific* models. Generalized Estimating Equations ([GEEs]) are a well-known computational tool for *marginal models*. With GEEs the marginal mean $\mu = E[\mathbf{Y}] = g^{-1}(\mathbf{X}\beta)$ should be correctly specified, in combination with a working assumption about the dependence between outcomes. Even though this working assumption may be wrong, the GEE estimator of β has nice properties (consistency, asymptotic normality with mean β and covariance matrix as in Liang and Zeger (1986)). Applications of GEEs in actuarial predictive modeling are in Purcaru et al. (2004) and Denuit et al. (2007). With this marginal modeling technique interest is only in the effect of the regression parameters on the marginal response; no cluster-specific inference or prediction is possible. Within the class of *conditional models* $\mathbf{Y}$ is modeled conditional upon (a subset of) the other outcomes. Our focus - from Section 3.3 on - is on *subject or cluster-specific models*, more specifically on generalized linear mixed models ([GLMMs]) where random, subject or cluster-specific effects are included in the linear predictor.

3.3 GLMMs: model specifications and assumptions

GLMMs extend GLMs by adding random effects $\mathbf{Z}\mathbf{u}$ to the linear predictor $\mathbf{X}\boldsymbol{\beta}$. Motivations for this extension are as stated in Section 1: the random effects enable cluster-specific prediction, they allow for heterogeneity between clusters and structure correlation within clusters. Conditionally on a q -dimensional vector $\mathbf{u}_i$ of random effects for cluster i , GLMM assumptions for the j th response on cluster i , Y_{ij} , are

$$\begin{aligned} Y_{ij}|\mathbf{u}_i &\sim f_{Y_{ij}|\mathbf{u}_i}(y_{ij}|\mathbf{u}_i) \\ f_{Y_{ij}|\mathbf{u}_i}(y_{ij}|\mathbf{u}_i) &= \exp\left(\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} - c(y_{ij}, \phi)\right). \end{aligned} \quad (50)$$

The following conditional relations hold

$$\mu_{ij} = E[Y_{ij}|\mathbf{u}_i] = b'(\theta_{ij}) \quad \text{and} \quad \text{Var}[Y_{ij}|\mathbf{u}_i] = \phi b''(\theta_{ij}) = \phi V(\mu_{ij}). \quad (51)$$

A transformation of the mean μ_{ij} is linear in both fixed and random effects

$$g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i, \quad (52)$$

with $\boldsymbol{\beta}$ the fixed effects parameter vector, $\mathbf{u}_i$ the vector of random effects for cluster i , $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ p and q dimensional vectors of known covariates corresponding with the fixed and random effects, respectively. A distributional assumption for the random effects vector $\mathbf{u}_i$, say $f_U(\mathbf{u}_i)$, completes the specification of a GLMM. Most applications use normally distributed random effects. XXX and XXX discuss examples with a different distributional assumption for the random effects.

The model assumptions in (50), (51) and (52) imply the following specifications for marginal mean and variance

$$\begin{aligned} E[Y_{ij}] &= E[E[Y_{ij}|\mathbf{u}_i]] = E[g^{-1}(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i)] \\ \text{Var}(Y_{ij}) &= \text{Var}(E[Y_{ij}|\mathbf{u}_i]) + E[\text{Var}(Y_{ij}|\mathbf{u}_i)] \\ &= \text{Var}(\mu_{ij}) + E[\phi V(\mu_{ij})] \\ &= \text{Var}(g^{-1}[\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i]) + E[\phi V(g^{-1}[\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i])]. \end{aligned} \quad (53)$$

In general, simplification of these expressions is not possible. The GLMM regression parameters $\boldsymbol{\beta}$ do not have a marginal interpretation; they express the effect of a set of covariates on the response, conditional on the random effects $\mathbf{u}_i$. Indeed, $E[Y_{ij}] = E[E[Y_{ij}|\mathbf{u}_i]] = E[g^{-1}(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_i)] \neq g^{-1}(\mathbf{x}_{ij}'\boldsymbol{\beta})$. Illustration 3.1 shows explicit calculation of marginal mean, variance and covariance within a Poisson GLMM.

Illustration 3.1 (A Poisson GLMM) Conditional on a random intercept $u_i \sim N(0, \sigma^2)$,

Y_{ij} is Poisson distributed with $\mu_{ij} = E[Y_{ij}|u_i] = \exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_i)$. Thus, the link function g is the logarithm. Straightforward calculations using mean and variance of a lognormal distribution show

$$\begin{aligned} E(Y_{ij}) &= E(E(Y_{ij}|u_i)) = E(\exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_i)) \\ &= \exp(\mathbf{x}_{ij}'\boldsymbol{\beta}) \exp(\sigma^2/2) \end{aligned} \quad (54)$$

and

$$\begin{aligned} \text{Var}(Y_{ij}) &= \text{Var}(E(Y_{ij}|u_i)) + E(\text{Var}(Y_{ij}|u_i)) \\ &= E(Y_{ij})(\exp(\mathbf{x}_{ij}'\boldsymbol{\beta})[\exp(3\sigma^2/2) - \exp(\sigma^2/2)] + 1), \end{aligned} \quad (55)$$

and

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(E(Y_{ij}|u_i), E(Y_{ik}|u_i)) + E(\text{Cov}(Y_{ij}, Y_{ik}|u_i)) \quad (j \neq k) \\ &= \exp(\mathbf{x}_{ij}'\boldsymbol{\beta}) \exp(\mathbf{x}_{ik}'\boldsymbol{\beta}) (\exp(2\sigma^2) - \exp(\sigma^2)). \end{aligned} \quad (56)$$

The expression in round parentheses in (55) is always bigger than 1. Thus, although $Y_{ij}|u_i$ follows a regular Poisson distribution, the marginal distribution of Y_{ij} is over-dispersed. According to (56), due to the random intercept, observations on the same subject are no longer independent. Actuarial literature on ratemaking (see e.g. Denuit et al. (2007) and Antonio and Valdez (2012)) often uses a slightly modified version of the normality assumption, namely $u_i \sim N(-\frac{\sigma^2}{2}, \sigma^2)$. This leads to

$$\begin{aligned} E[Y_{ij}] &= E[E(Y_{ij}|u_i)] = \exp(\mathbf{x}_i'\boldsymbol{\beta} - \frac{\sigma^2}{2} + \frac{\sigma^2}{2}) \\ &= \exp(\mathbf{x}_i'\boldsymbol{\beta}), \\ E[Y_i|u_i] &= \exp(\mathbf{x}_i'\boldsymbol{\beta} + u_i). \end{aligned} \quad (57)$$

In actuarial parlance, the so-called **a priori** premium, specified as $\exp(\mathbf{x}_i'\boldsymbol{\beta})$, uses only a priori measurable risk factors (like gender, age, car capacity, ...). It is the marginal mean of Y_{ij} and is therefore correct on average. The **a posteriori** correction factor, $\exp(u_i)$, adjusts the a priori tariff based on the observed claim history of the insured (through prediction of u_i or by using its posterior distribution).

Illustration 3.2 (An illustration of shrinking) We consider a claim frequency model using the auto claim data from Yip and Yau (2005), where we specify a log-linear Poisson model with *JOBCLASS* as random effects. In particular, we are interested at how the estimate for each job class level differs between the mixed model and the GLM where *JOBCLASS* enters as a categorical fixed effect. Figure 6 shows such a comparison on the estimation of job class levels. The horizontal dotted line corresponds to the estimated intercept from the mixed model and represents the average effect for all job categories because all the random effects have zero means. That is, it is roughly the estimate

when all job categories are pooled together. On the other hand, the estimates from the generalized linear model (the red points) can be viewed as the individual estimate for each job class level ignoring the other levels - indeed, fitting a GLM with only the job class as a predictor is equivalent to fitting 8 separate GLMs on each subset of data with a unique job class because of the orthogonal design matrix corresponding to the job class. We see that the mixed model (the green triangle) shrinks the separate estimates from the GLM toward the pooled group-level estimate across all the job classes. The shrinkage is most significant for Lawyer, Professional and Student. Therefore, the generalized linear mixed model captures the core insight of the credibility models, where the estimates from the mixed models are can be viewed as the weighted average between the pooled group-level estimate and the separate individual estimates. As a result, the mixed model produces less extreme estimates while still accounting for the heterogeneity across the various levels.

Figure 6: The job class estimates from the generalized linear model (●) and the Poisson mixed models (△) in the auto insurance frequency model. The horizontal line is the average estimate for all job classes, and the vertical lines show the uncertainty intervals based on $\pm$ one standard errors.

3.4 GLMMs: estimation

Using the model specifications in (50) it is straightforward to write down the likelihood of the corresponding GLMM

$$L = \int f_{Y|u}(y|u)f_U(u)du, \quad (58)$$

where the integral goes over the q dimensional random effects vector $\mathbf{u}$. For ease of explanation we will start from a basic example. Let Y_{ij} denote the number of claims registered on policy i in period j , and model Y_{ij} with a Poisson GLMM with random intercept

$$\begin{aligned} Y_{ij} | u_i &\sim \text{indep. Poi}(\mu_{ij}) \\ \log \mu_{ij} &= \mathbf{x}'_{ij} \boldsymbol{\beta} + u_i \\ u_i &\sim \text{i.i.d. } N(0, \sigma^2) \\ L &= \prod_{i=1}^m \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} \frac{\mu_{ij} e^{-\mu_{ij}}}{y_{ij}!} \right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} u_i^2} du_i. \end{aligned} \quad (59)$$

The presence of the integral in (59) hinders maximum likelihood estimation and prohibits explicit expressions for estimators and predictors, like those derived in (26) and (27) for LMMs. Only so-called *conjugate* distributional specifications lead to a closed-form solution in (59); a normal distribution for the response, combined with normally distributed random effects (as with LMMs) being one example. More general model assumptions require approximative techniques to estimate $\boldsymbol{\beta}$, $\mathbf{D}$ and predict the random effect for cluster i , u_i . As in Molenberghs and Verbeke (2005) we distinguish three approaches to tackle this problem: approximating the integrand, approximating the data and approximating the integral (through numerical integration). Having Pinheiro and Bates (2000), McCulloch and Searle (2001) (Chapters 8 and 10) and Tuerlinckx et al. (2006) as main references, we discuss below some highlights of these methods. Section 3.4.5 presents pros and cons of the techniques mentioned in 3.4.1, ?? and 3.4.3, as well as references to some other methods (not discussed here). We postpone a discussion of Bayesian methods to Section 5.

3.4.1 Approximating the likelihood: the Laplace method

The Laplace method (see Tierny and Kadane (1986)) approximates integrals of the form

$$\int e^{h(\mathbf{u})} d\mathbf{u}, \quad (60)$$

with h a smooth, bounded and unimodal function of a q -dimensional variable $\mathbf{u}$. The method relies on a second-order Taylor expansion of $h(\mathbf{u})$ around its maximum $\hat{\mathbf{u}}$

$$h(\mathbf{u}) \approx h(\hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}})' h''(\hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}}), \quad (61)$$

with

$$\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} \Big|_{\mathbf{u}=\hat{\mathbf{u}}} = \mathbf{0}, \quad (62)$$

and $h''(\hat{\mathbf{u}}) = \frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} \big|_{\mathbf{u}=\hat{\mathbf{u}}}$ the matrix with second order derivatives of h , evaluated at $\hat{\mathbf{u}}$. We replace $h(\mathbf{u})$ with the approximation from (61)

$$\int e^{h(\mathbf{u})} d\mathbf{u} \approx \int e^{h(\hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u}-\hat{\mathbf{u}})' h''(\hat{\mathbf{u}})(\mathbf{u}-\hat{\mathbf{u}})} d\mathbf{u}. \quad (63)$$

Approximating the density of $\mathbf{u}$ with a multivariate Gaussian distribution $\mathcal{N}(\hat{\mathbf{u}}, (-h''(\hat{\mathbf{u}}))^{-1})$ leads to

$$\int e^{h(\mathbf{u})} d\mathbf{u} \approx (2\pi)^{q/2} | -h''(\hat{\mathbf{u}}) |^{-1/2} e^{h(\hat{\mathbf{u}})}. \quad (64)$$

This technique is readily available to approximate the likelihood in a GLMM (see Breslow and Clayton (1993) and McCulloch and Searle (2001), among other references)

$$\begin{aligned} \ell &= \log \int f_{Y|U}(\mathbf{y}|\mathbf{u}) f_U(\mathbf{u}) d\mathbf{u} \\ &= \log \int e^{\log f_{Y|U}(\mathbf{y}|\mathbf{u}) + \log f_U(\mathbf{u})} d\mathbf{u} \\ &= \log \int e^{h(\mathbf{u})} d\mathbf{u}, \end{aligned} \quad (65)$$

with $h(\mathbf{u}) := \log f_{Y|U}(\mathbf{y}|\mathbf{u}) + \log f_U(\mathbf{u}) = \log f_{Y|U}(\mathbf{y}|\mathbf{u}) - \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u} - \frac{q}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{D}|$. (62) should be solved numerically and requires

$$\begin{aligned} \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} &= \frac{\partial \log f_{Y|U}(\mathbf{y}|\mathbf{u})}{\partial \mathbf{u}} - \mathbf{D}^{-1} \mathbf{u} = \mathbf{0} \\ &\quad \Updownarrow \\ \frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{D}^{-1} \mathbf{u} &= \mathbf{0}, \end{aligned} \quad (66)$$

where $\mathbf{W}$ and Δ are diagonal matrices with elements $[V(\mu_i)(g'(\mu_i))^2]^{-1}$ and $g'(\mu_i)$, respectively⁷. Hereby $g(\mu_i)$ and $V(\mu_i)$ are the conditional mean and variance of Y_i , given $\mathbf{u}$, as introduced in (48).

We also need the matrix of second order derivatives (see (64))

$$\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} = -\frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta \frac{\partial \boldsymbol{\mu}}{\partial \mathbf{u}'} + \frac{1}{\phi} \mathbf{Z}' \frac{\partial \mathbf{W} \Delta}{\partial \mathbf{u}'} (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{D}^{-1}. \quad (67)$$

⁷Derivations are similar to those in Chapter XXX on GLMs, and basically go as follows:

$$\begin{aligned} \frac{\partial \log f_{Y|U}(\mathbf{y}|\mathbf{u})}{\partial \mathbf{u}} &= \frac{1}{\phi} \sum_i \left(y_i \frac{\partial \theta_i}{\partial \mathbf{u}} - \frac{\partial \psi(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mathbf{u}} \right) \\ &= \frac{1}{\phi} \sum_i (y_i - \mu_i) \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} \mathbf{z}_i'. \end{aligned}$$

The random vector corresponding with the second term in this expression has expectation zero, with respect to $f_{Y|U}(\mathbf{y}|\mathbf{u})$, and will be ignored. Therefore,

$$\begin{aligned} -\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} &= \frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta \Delta^{-1} \mathbf{Z} + \mathbf{D}^{-1} \\ &= \left(\frac{1}{\phi} \mathbf{Z}' \mathbf{W} \mathbf{Z} \mathbf{D} + \mathbf{I} \right) \mathbf{D}^{-1}. \end{aligned} \quad (68)$$

Using this expression an approximation to the log-likelihood in (65) follows

$$\begin{aligned} \ell &\approx \log f_{Y|U}(\mathbf{y}|\hat{\mathbf{u}}) - \frac{1}{2} \hat{\mathbf{u}}' \mathbf{D}^{-1} \hat{\mathbf{u}} - \frac{q}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{D}| \\ &\quad + \frac{q}{2} \log 2\pi - \frac{1}{2} \log |(\mathbf{Z}' \mathbf{W} \mathbf{Z} \mathbf{D} / \phi + \mathbf{I}) \mathbf{D}^{-1}| \\ &= \log f_{Y|U}(\mathbf{y}|\hat{\mathbf{u}}) - \frac{1}{2} \hat{\mathbf{u}}' \mathbf{D}^{-1} \hat{\mathbf{u}} + \frac{1}{2} \log |\mathbf{Z}' \mathbf{W} \mathbf{Z} \mathbf{D} / \phi + \mathbf{I}|. \end{aligned} \quad (69)$$

This expression should be maximized with respect to β . Assuming $\mathbf{W}$ is not changing a lot as a function of β , the last term can be ignored and

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\phi} \mathbf{X}' \mathbf{W} \Delta (\mathbf{y} - \mu). \quad (70)$$

Therefore, the following set of equations has to be solved simultaneously with respect to β and $\mathbf{u}$ (using a numerical optimization method)

$$\begin{aligned} \frac{1}{\phi} \mathbf{X}' \mathbf{W} \Delta (\mathbf{y} - \mu) &= \mathbf{0} \\ \frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta (\mathbf{y} - \mu) &= \mathbf{D}^{-1} \mathbf{u}. \end{aligned} \quad (71)$$

This set of equations also arises by jointly maximizing (with respect to β and $\mathbf{u}$)

$$\log f_{Y|U}(\mathbf{y}|\mathbf{u}) - \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u}, \quad (72)$$

which is a quasi-likelihood term, $f_{Y|u}(\mathbf{y}|\mathbf{u})$, augmented with a penalty term, $\mathbf{u}' \mathbf{D} \mathbf{u}$. Hence, the name Penalized Quasi-Likelihood (PQL) for (72). Breslow and Clayton (1993) present a Fisher scoring algorithm, and its connection with Henderson's mixed model equations (see (32)), for simultaneous solution of the set of equations in (71).

3.4.2 Approximating the data: pseudo-likelihood (PL)

Wolfinger and O'Connell (1993) develop pseudo-likelihood ([PL]) (or restricted pseudo-likelihood, [REPL]) in the context of GLMMs. This approach generalizes the idea of a

working variate, introduced for MLE with GLMs (see Chapter XXX), to the case of GLMMs (also see Breslow and Clayton (1993) and McCulloch and Searle (2001)). In the context of GLMs Nelder and Wedderburn (1972) define a working variate t_i as follows

$$\begin{aligned} t_i &= g(\mu_i) + g'(\mu_i)(Y_i - \mu_i) \\ &= \mathbf{x}_i' \boldsymbol{\beta} + g'(\mu_i)(Y_i - \mu_i). \end{aligned} \quad (73)$$

Estimates of $\boldsymbol{\beta}$ follow from iteratively fitting a weighted linear regression of $\mathbf{t}$ on $\mathbf{X}$, until convergence of the estimates. In a GLMM we generalize the notion of a working variate t_i as follows

$$t_i = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \mathbf{u} + g'(\mu_i)(Y_i - \mu_i). \quad (74)$$

This is a first order Taylor expansion of $g(y_i)$ around μ_i . In matrix notation the vector of working variates, $\mathbf{t}$, becomes

$$\mathbf{t} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu}), \quad (75)$$

with $\boldsymbol{\Delta}$ a diagonal matrix with entries $g'(\mu_i)$. Calculating the variance of $\mathbf{t}$ is complicated because of the dependence of $\boldsymbol{\Delta}$ on $\boldsymbol{\mu}$ (and therefore on the random vector $\mathbf{u}$). A simplification is possible by replacing $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}$ in the variance matrix (see Wolfinger and O'Connell (1993)). Consequently,

$$\begin{aligned} \text{Var}(\mathbf{t}) &= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \boldsymbol{\Delta}_{\hat{\boldsymbol{\mu}}} \text{Var}(\mathbf{Y} - \boldsymbol{\mu})_{\hat{\boldsymbol{\mu}}} \boldsymbol{\Delta}_{\hat{\boldsymbol{\mu}}} \\ &:= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \boldsymbol{\Sigma}_{\hat{\boldsymbol{\mu}}}. \end{aligned} \quad (76)$$

The working variate $\mathbf{t}$ approximately follows a linear mixed model (as in (14)), with design matrices $\mathbf{X}$ (fixed effects), $\mathbf{Z}$ (random effects) and $\mathbf{D}$ the covariance matrix of the distribution of random effects. In this LMM it is straightforward to estimate $\boldsymbol{\beta}$, $\mathbf{u}$ and the unknown variance components. Therefore, the pseudo-likelihood algorithm goes as follows. Starting from initial estimates of $\boldsymbol{\beta}$, $\mathbf{u}$ and the variance components, the working variates in (75) are evaluated. Consequently, using LMM methodology, updated estimates follow from (75) and (76). These steps are repeated until convergence of the estimates.

3.4.3 Approximating the integral: numerical integration techniques

Approximating the integral in (58) with a so-called (adaptive) quadrature rule for numerical integration is based upon Liu and Pierce (1994). For ease of explanation we consider below the case of a one-dimensional integral. The case with multidimen-

sional integrals is documented in Tuerlinckx et al. (2006).

Non-adaptive Gauss–Hermite quadrature. *Non-adaptive* Gauss-Hermite quadrature approximates an integral of the form

$$\int_{-\infty}^{+\infty} h(z) \exp(-z^2) dz, \quad (77)$$

with a weighted sum, namely

$$\int_{-\infty}^{+\infty} h(z) \exp(-z^2) dz \approx \sum_{l=1}^Q w_l h(z_l). \quad (78)$$

Q denotes the order of the approximation, the z_l are the zeros of the Q th order Hermite polynomial and the w_l are corresponding weights. The nodes (or quadrature points) z_l and the weights w_l are tabulated in Abramowitz and Stegun (1972) (page 924). The quadrature points used in (78) do not depend on h . As such, it is possible that only very few nodes lie in the region where most of the mass of h is, which would lead to poor approximations.

Adaptive Gauss–Hermite quadrature. With an *adaptive* Gauss-Hermite quadrature rule the nodes are rescaled and shifted such that the integrand is sampled in a suitable range. Assume $h(z)\phi(z;0,1)$ is unimodal and consider the numerical integration of $\int_{-\infty}^{+\infty} h(z)\phi(z;0,1)dz$. Let $\hat{\mu}$ and $\hat{\nu}$ be

$$\hat{\mu} = \text{mode}[h(z)\phi(z;0,1)] \quad \text{and} \quad \hat{\nu}^2 = \left[-\frac{\partial^2}{\partial z^2} \ln(h(z)\phi(z;0,1)) \Big|_{z=\hat{\mu}} \right]^{-1}. \quad (79)$$

Acting as if $h(z)\phi(z;0,1)$ were a Gaussian density, $\hat{\mu}$ and $\hat{\nu}$ would be the mean and variance of this density. The quadrature points in the adaptive procedure, z_l^* , are centered at $\hat{\mu}$ with spread determined by $\hat{\nu}$, namely

$$z_l^* = \hat{\mu} + \sqrt{2}\hat{\nu}z_l \quad (80)$$

with $(l = 1, \dots, Q)$. Now rewrite $\int_{-\infty}^{+\infty} h(z)\phi(z;0,1)dz$ as

$$\int_{-\infty}^{+\infty} \frac{h(z)\phi(z;0,1)}{\phi(z;\mu,\nu)} \phi(z;\mu,\nu) dz, \quad (81)$$

where $\phi(z; \mu, \nu)$ is the Gaussian density function with mean μ and variance ν^2 . Using simple manipulations it is easy to see that for a suitably regular function v

$$\begin{aligned} \int_{-\infty}^{+\infty} v(z) \phi(z; \mu, \nu) dz &= \int_{-\infty}^{+\infty} v(z) (2\pi\nu^2)^{-1/2} \exp\left(-\frac{1}{2} \left(\frac{z-\mu}{\nu}\right)^2\right) dz \\ &= \int_{-\infty}^{+\infty} \frac{v(\mu + \sqrt{2}\nu z)}{\sqrt{\pi}} \exp(-z^2) dz \\ &\approx \sum_{l=1}^Q \frac{v(\mu + \sqrt{2}\nu z_l)}{\sqrt{\pi}} w_l. \end{aligned} \quad (82)$$

Using $\frac{h(z)\phi(z;0,1)}{\phi(z;\mu,\nu)}$ instead of $v(z)$ and replacing μ and ν with their estimates from (79), results in the following quadrature formula

$$\begin{aligned} \int_{-\infty}^{+\infty} h(z) \phi(z; 0, 1) dz &\approx \sqrt{2}\hat{\nu} \sum_{l=1}^Q w_l \exp(z_l^2) \phi(z_l^*; 0, 1) h(z_l^*) \\ &= \sum_{l=1}^Q w_l^* h(z_l^*), \end{aligned} \quad (83)$$

with adaptive weights $w_l^* := \sqrt{2}\hat{\nu} w_l \exp(z_l^2) \phi(z_l^*; 0, 1)$. (83) is an *adaptive* Gauss-Hermite quadrature formula.

Link with Laplace approximation. We illustrate the connection between the Laplace approximation (from Section 3.4.1) and adaptive Gauss-Hermite quadrature with a single node. Indeed, when $Q = 1$ (i.e. the case of a single node), $z_1 = 0$ (from the Hermite polynomial) and $w_1 = 1$. The corresponding adaptive node and weight are $z_1^* = \hat{\mu}$ and $w_1^* = \sqrt{2}\hat{\nu} \phi(z_1^*; 0, 1)$. The adaptive GH quadrature formula then becomes

$$\begin{aligned} \int h(z) \phi(z; 0, 1) dz &\approx \sqrt{2}\hat{\nu} \exp(\log(\phi(z_1^*; 0, 1) h(z_1^*))) \\ &\propto (2\pi)^{1/2} \underbrace{\left| -\frac{\partial^2}{\partial z^2} \log(h(z) \phi(z; 0, 1)) \right|_{z=\hat{\mu}} }_{\hat{\nu}}^{-1/2} \exp\{\log(\phi(\hat{\mu}; 0, 1) h(\hat{\mu}))\}, \end{aligned} \quad (84)$$

where $\hat{\mu} = z_1^*$ maximizes $h(z) \phi(z; 0, 1)$. This corresponds with the Laplace formula from (64).

Adaptive Gauss-Hermite quadrature for GLMMs. We describe the case of a GLMM with a single, normally distributed random effect $u_i \sim N(0, \sigma)$ for each cluster i . The use of adaptive Gauss-Hermite quadrature with GLMMs starts from determining the

posterior mode of u_i . Since this posterior distribution depends on unknown fixed effects and variance parameters, we replace the unknown β , ϕ and σ with their current estimates: $\hat{\beta}^{(c)}$, $\hat{\phi}^{(c)}$ and $\hat{\sigma}^{(c)}$. Using these current estimates $\hat{u}_i$ maximizes

$$f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}), \quad (85)$$

which is proportional to the posterior density of u_i , given $\mathbf{y}_i$

$$\begin{aligned} f(u_i|\mathbf{y}_i) &= \frac{f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})}{\int f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})du_i} \\ &\propto f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}). \end{aligned} \quad (86)$$

Therefore $\hat{u}_i$ is the posterior mode of u_i . We also determine (numerically) $\hat{v}_i^2$ as

$$\hat{v}_i^2 = \left[-\frac{\partial^2}{\partial u_i^2} \ln(f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})) \Big|_{u_i=\hat{u}_i} \right]^{-1}. \quad (87)$$

Using an adaptive Gauss–Hermite quadrature rule we approximate the likelihood contribution of cluster i as follows (with $\delta_i := \sigma^{-1}u_i \sim N(0, 1)$)

$$\begin{aligned} \int f_{\mathbf{Y}|U}(\mathbf{y}_i|u_i)f_U(u_i)du_i &= \int f_{\mathbf{Y}|U}(\mathbf{y}_i|\delta_i)\phi(\delta_i|0, 1)d\delta_i \\ &= \int \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|\delta_i) \right) \phi(\delta_i|0, 1)d\delta_i \\ &= \int \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|\delta_i) \right) \frac{\phi(\delta_i|0, 1)}{\phi(\delta_i|\hat{\delta}_i, \hat{v}_i)} \phi(\delta_i|\hat{\delta}_i, \hat{v}_i)d\delta_i \\ &\approx \sum_{l=1}^Q w_l^* \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|z_l^*) \right), \end{aligned} \quad (88)$$

with adaptive weights $w_l^* = \sqrt{2}\hat{v}_i w_l \exp(z_l^2)\phi(z_l^*; 0, 1)$ and $z_l^* = \hat{\delta}_i + \sqrt{2}\hat{v}_i z_l$. In this expression the linear predictor corresponding with $f_{Y|U}(y_{ij}|\delta_i)$ and $f_{Y|U}(y_{ij}|z_l^*)$, respectively, is $\mathbf{x}_{ij}'\beta + \sigma\delta_i$ and $\mathbf{x}_{ij}'\beta + \sigma z_l^*$. Multiplying (88) over all clusters i leads to the total likelihood. Maximizing the latter over the fixed effects regression parameters, the dispersion parameter and the variance components leads to updated parameter estimates $\hat{\beta}^{(c+1)}$, $\hat{\phi}^{(c+1)}$ and $\hat{\sigma}^{(c+1)}$. We predict the cluster-specific random effects with the posterior modes using (85).

3.4.4 More on estimation with GLMMs: Tweedie compound Poisson mixed models

For the Tweedie compound Poisson mixed models (I assume that this was introduced in previous chapters) with a variance function $V(\mu) = \mu^p$ for some $p \in (1, 2)$, we seek to estimate the unknown variance function, i.e., the index parameter p from the data along with the fixed effects and the variance component. This parameter has a significant impact on hypothesis tests and predictive uncertainty measures (Davidian and Carroll, 1987; Peters et al., 2009; Zhang, 2012), which is of special interest to the insurance industry. For example, if a Tweedie compound Poisson GLM is exploited in loss reserving modeling, the uncertainty measures of the predicted outstanding liability will be substantially influenced by the choice/estimation of the index parameter.

One approach in estimating the variance function is using the profile likelihood (Cox and Reid, 1987). For the compound Poisson distribution, such an approach must be implemented based on the true likelihood rather than the quasi-likelihood. It is well known that the basic quasi-likelihood method, and hence the PQL method introduced above is not equipped to estimate the unknown variance function. Its natural extension, the extended quasi-likelihood (Nelder and Pregibon, 1987), however, is also not well suited to this task in that it involves a term $\log(V(y))$ which becomes infinite for $y = 0$. Its implementation therefore requires adding a small positive constant to the observed zeros which, unfortunately, is highly influential on parameter estimation (see Zhang (2012)).

Likelihood-based methods, namely, the Laplace approximation and the adaptive Gauss-Hermite quadrature method, must be used to enable data-driven estimation of the index parameter. Yet, a complicating factor is that the compound Poisson distribution has an intractable density function. When performing maximum likelihood estimation, we must rely on numerical methods to approximate the density function, that is, the conditional distribution of the data given the random effects. Such numerical methods that allow fast and accurate evaluation of the compound Poisson density function are provided in Dunn and Smyth (2005, 2008). Similarly to the above, the approximated likelihood is then optimized numerically to produce parameter estimates, including the maximum likelihood estimate for p .

3.4.5 Pros and cons of various estimation methods for GLMMs

Laplace and PQL methods for estimation within GLMMs rely on quite a few approximations. Breslow and Lin (1995) and Lin and Breslow (1996) investigate settings in which PQL performs poorly, and discuss the limits of this approach. Based on this McCulloch and Searle (2001) decide “*We thus cannot recommend the use of simple PQL methods in practice.*” (see McCulloch and Searle (2001), Chapter 10, page 283). Gauss–

Hermite quadrature is more accurate than PQL but limited to GLMMs with a small number of nested random effects. It is not possible to handle a large number of random effects, crossed random effects or high levels of nesting with this approach. Moreover, Gauss–Hermite quadrature is explicitly designed for normally distributed random effects, although other quadrature formulas exist (not discussed here).

The (Monte Carlo) EM algorithm and simulated maximum likelihood or Monte Carlo integration (see McCulloch and Searle (2001), Chapter 10, or Tuerlinckx et al. (2006)) are alternative methods for estimation with GLMMs.

We discuss a Bayesian implementation of (G)LMMs in Section 5. This is a way to circumvent the estimation problems discussed above.

Illustration 3.3 (A GLMM for claim frequencies) *We analyze the data from Illustration 1.2. The data are claim counts and our interest lies in predictions regarding individual risk classes. We therefore explore the use of a Poisson GLMM. Our model specifications are*

$$\text{Count}_{ij} | \mathbf{u}_i \sim \text{Poisson}(\mu_{ij})$$

$$\text{with } \log(\mu_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0 + \beta_1 \text{Year}_{ij} + u_{i,0}, \quad (89)$$

$$\text{and } \log(\mu_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0 + \beta_1 \text{Year}_{ij} + u_{i,0} + u_{i,1} \text{Year}_{ij}. \quad (90)$$

Count_{ij} is the number of claims in year j , reported for subject or risk class i . β_0 and β_1 are fixed effects, $u_{i,0}$ and $u_{i,1}$ denote a risk class specific intercept, respectively slope. We assume: $\mathbf{u}_i = (u_{i,0}, u_{i,1})' \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$. Across subjects, random effects are independent. To enable out-of-sample predictions, we split the data set in a training (without Count_{i7}) versus validation set (the Count_{i7} observations). Models are estimated on the training set, and centering of Year is applied. Tables 1 and 2 use the following notation: $\theta_0 = \text{Var}(u_{i,0})$, $\theta_1 = \text{Var}(u_{i,1})$ and $\theta_{0,1} = \theta_{1,0} = \text{Cov}(u_{i,0}, u_{i,1})$.

The table shows parameter estimates (and standard errors, at least for fixed effects⁸) as obtained with standard routines in R and SAS. Data and programming code for this example are available in XXX. In R we use the `glmer` routine from library `lme4` (for Laplace approximation and adaptive Gauss–Hermite quadrature). `glmmPQL` from library `MASS` estimates the GLMM by repeatedly fitting a linear mixed model (with `lme` from `nlme`) to an approximation of the data. Laplace approximation and pseudo-likelihood ((RE)PL) is available in SAS through `Proc Glimmix`. For adaptive Gauss–Hermite quadrature we use `Proc NLMixed`. The programming code illustrates how estimates for fixed and random effects, as well as variance components, can be retrieved. Note that, in Table 2, results obtained with PL are missing, due to lack of convergence of the routines in SAS and R. Illustration 5.1 puts focus on prediction of future observa-

[Reference to availability of programming code.]

⁸For reasons discussed in XXX (reference to Bates) R routines do not report s.e. corresponding with variance components. This is in contrast with standard output from SAS procedures.

tions for specific risk classes and compares the results of a Bayesian analysis with the results obtained here.

Tables 1 (for model (89)) and 2 (for model (90)) show that parameter estimates obtained with different methods, and different implementations, are close. The optimal value of the log-likelihood, and corresponding AIC, is not printed for PL, since this method works on a likelihood of approximated data (as explained in Section 3.4.2), and not on the true likelihood. In the second column, the optimal log-likelihood for model (89) is -883. With 2 parameters from the fixed effects and 1 variance parameter, this results in an AIC of: $(-2) \cdot (-883) + 2 \cdot 3 = 1,771$. Comparing model (89) with (90) results in an observed LRT statistic of $(-2)((-883) - (-845.7)) = 73.8$. We therefore prefer the model with varying intercepts and slopes.

	Laplace (Section 3.4.1)		PL (Section 3.4.2)		adaptive G-H (Section 3.4.3)	
	R	SAS	R	SAS	R	SAS
Model (89)						
β_0	-3.562 (0.0835)	-3.562 (0.0838)	-3.49 (0.0891)	-3.533 (0.0827)	-3.562 (0.0836)	-3.562 (0.0839)
β_1	0.00973 (0.00516)	0.00973 (0.00516)	0.0095 (0.00656)	0.00973 (0.00516)	0.00973 (0.00516)	0.00973 (0.00516)
θ_0	0.805	0.805 (0.114)	0.863	0.79 (0.11)	0.807	0.807 (0.114)
logLik	-883	-2,197	/	/	-882	-2,197
AIC	1,772	4,401	/	/	1,771	4,400

Table 1: Estimation results corresponding with (89), workers' compensation insurance (frequencies) introduced in Illustration 1.2.

3.5 Statistical inference with GLMMs

The general ideas on statistical inference with LMMs carry over to GLMMs where fitting is based on maximum likelihood principles. Wald, score and likelihood ratio tests are available for testing fixed effects parameters, as well as variance components. However, closed-form expressions, for example for the covariance matrix of $\hat{\beta}$, are no longer available. Numerical evaluation of the inverse Fisher information matrix is required for precision estimates. When using the PL method as described in Section 3.4.2, the original likelihood expression should be used in a LRT, and not the likelihood of the LMM that is specified for the pseudo-data. As with LMMs, testing the necessity of a random effect is problematic, since the corresponding null hypothesis is on the boundary of the parameter space for variance components. With respect to inference

	<i>Laplace (Section 3.4.1)</i>		<i>adaptive G-H (Section 3.4.3)</i>	
	R	SAS	R	SAS
Model (90)				
β_0	-3.569 (0.0836)	-3.569 (0.0839)	-3.569 (0.0838)	-3.569 (0.0841)
β_1	0.00892 (0.0107)	0.00892 (0.0111)	0.00892 (0.0107)	0.00891 (0.0111)
θ_0	0.807	0.807 (0.114)	0.810	0.810 (0.115)
θ_1	0.00638	0.00638 (0.00174)	0.00639	0.00639 (0.00174)
$\theta_{0,1}$	-0.00222	-0.00223 (0.00174)	-0.00223	-0.00224 (0.0104)
logLik	-845.7	-2,160	-771.1	-2,160
AIC	1,701	4,331	1,552	4,321

Table 2: Estimation results corresponding with (90), workers' compensation insurance (frequencies) introduced in Illustration 1.2.

with (G)LMMs a Bayesian analysis has some additional features, see Section 5 for discussion.

4 Non-linear mixed models

LMMs and GLMMs model the mean (in LMMs) or a transformation of the (conditional) mean (in GLMMs) as *linear* in the fixed effects parameters β and the random effects $\mathbf{u}$. Non-linear mixed models ([NLMM]) release the concept of linear predictors. In a NLMM the conditional distribution of Y_{ij} (being the j th response on cluster i), given $\mathbf{u}_i$, belongs to the exponential family with mean structure

$$E[Y_{ij}|\mathbf{u}_i] = h(\mathbf{x}_{ij}, \beta, \mathbf{z}_{ij}, \mathbf{u}_i), \quad (91)$$

where $h(\cdot)$ is an arbitrary function of covariates, parameters and random effects. A distributional assumption for the random effects completes the model assumptions; typically $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{D})$. GLMMs are therefore a subclass of the general class of NLMMs. (Adaptive) Gauss–Hermite quadrature is available for ML estimation within NLMMs. A fully Bayesian analysis is an alternative approach.

Illustration 4.1 (A NLMM for loss reserving) *Motivated by accounting with-in cohort*

dependency arising from longitudinal data and inspired by detecting the patterns in Figure 3, where moderate variations of loss growth across accident years as well as companies are observed, we construct a nonlinear hierarchical model that enable us to carry out inference at the level of industry, company and accident years.

We will use $y_{ik}(t_j)$ to denote the cumulative loss of accident year i and company k at the j_{th} evaluation. To reflect the longitudinal nature of the loss payment amounts, we specify the model on the basis of cumulative losses instead of incremental losses as follows:

$$\begin{aligned}\log y_{ik}(t_j) &= \log \mu_{ik}(t_j) + \epsilon_{ik}(t_j), \\ \mu_{ik}(t_j) &= p_{ik} \cdot \gamma_{ik} \cdot G(t_j; \Theta_k).\end{aligned}\tag{92}$$

In the above equation, p_{ik} and γ_{ik} are the (given) premium and the (unknown) expected ultimate loss ratio for the i_{th} accident year in company k . Therefore $p_{ik} \cdot \gamma_{ik}$ equals the expected ultimate loss for the i_{th} accident year. $G(t_j, \Theta_k)$ is a parametric growth curve that depends on parameters Θ_k and measures the percentage of ultimate losses that have emerged as of time t_j . Therefore G must have the properties that $G(t_0; \Theta_k) = 0$ and $G(t_j; \Theta_k) \rightarrow 1$ as $t_j \rightarrow t_\infty$, where t_∞ denotes the time point at which loss development terminates. Distribution of $\epsilon_{ik}(t_j)$ must be specified, e.g., Zhang et al. (2012) exploited an Normal distribution allowing autocorrelation along the accident years.

In (92), the growth curve $G(t_j; \Theta_k)$ is chosen to be the log-logistic curve

$$G(t_j; \omega_k, \theta_k) = \frac{t_j^{\omega_k}}{t_j^{\omega_k} + \theta_k^{\omega_k}}.\tag{93}$$

To avoid overfitting the data, the model is specified in such a way as to allow the ultimate loss ratios γ_{ik} to vary by both company and accident year, while the growth parameters $(\omega_k, \theta_k)'$ are allowed only to vary by company (we specify the distributions on the logarithmic scale since these parameters must be positive):

$$\begin{aligned}\log \gamma_{ik} &\sim N(\log \gamma_k, \sigma_{\gamma_{year}}^2) \text{ for each year } i, \\ \log(\gamma_k, \omega_k, \theta_k)' &\sim N[\log(\gamma, \omega, \theta)', \Sigma] \text{ for each company } k,\end{aligned}\tag{94}$$

where $\sigma_{\gamma_{year}}$ is the accident-year level variation of the loss ratios on the logarithmic scale, and $(\gamma_k, \omega_k, \theta_k)'$ is the company level parameters.

We note that the nonlinear structure implicit in (92) and (93) brings several advantages in modeling claim triangle data. First, the nonlinear structure is intuitively appealing, as can be seen by examining Figure 3. Second, the expected ultimate loss $E[y_{ik}(t_\infty)] = p_{ik} \times \gamma_{ik}$ is explicitly modeled, and variability measure can therefore be

calculated directly from the model. Third, the ultimate loss ratio parameter is of fundamental importance in insurance company operations, and company management will typically possess relevant expert opinions and/or industry data pertaining to its estimation that can be naturally incorporated into one's analysis through a Bayesian prior. Finally, model predictions at any evaluation age can also be read directly from the growth curve.

5 Bayesian approach to (L,GL,NL)MMs

The presence of random effects is an essential feature in the hierarchical model formulation of a mixed model. A link with Bayesian statistics is then straightforward, since the random effects have explicit distributional assumptions. In addition to the distribution of the random effects $\mathbf{u}$ and the distributional framework for the response $\mathbf{Y}$, a Bayesian analysis requires prior distributions for β , (ϕ in GLMMs) and $\mathbf{D}$. Inference is based on simulated samples from the posterior distribution of the parameters, which is (with m clusters)

$$\begin{aligned} & f(\beta, \mathbf{D}, \phi, \mathbf{u}_1, \dots, \mathbf{u}_m | \mathbf{Y}_1, \dots, \mathbf{Y}_m) \\ \propto & \prod_{i=1}^m f_i(\mathbf{Y}_i | \beta, \phi, \mathbf{u}_1, \dots, \mathbf{u}_m) \prod_{i=1}^m f(\mathbf{u}_i | \mathbf{D}) f(\mathbf{D}) f(\beta) f(\phi). \end{aligned} \quad (95)$$

We refer to Chapters XXX and XXX on Bayesian concepts and regression models for an overview of useful concepts and simulation methods. For GLMMs in particular Zhao et al. (2006) and the references herein are a nice starting point for Bayesian (L,G,NL)MMs.

In light of our discussion in previous Sections, Bayesian multilevel models have some very nice features. As discussed in Section 2.3.1, precision estimates based on MLE require variance components estimates to be plugged in, and are therefore not able to account for all sources of randomness. A fully Bayesian approach, with a prior specified for each parameter (vector), solves this issue and provides a way to circumvent otherwise intractable calculations. The likelihood approximations discussed in Section 3.4 are replaced in a Bayesian analysis with general MCMC methodology for sampling from posterior distributions. This allows specification of more complex hierarchical model structures, such as the spatial structures in Chapter XXX or the 3-level count data models in Antonio et al. (2010). Moreover, the Bayesian methodology is not limited to Gaussian random effects and the use of non-Gaussian distributions can be explored as well. For predictive modeling in insurance Bayesian statistics is particularly useful for simulation from the posterior (predictive) distribution of quantities of interest, such as a policy's random effect or the number of claims in a future time

[Reference
to Chapters
on Bayesian
statistics.]

period.

Illustration 5.1 (A Bayesian GLMM) We revisit the data from Illustration 1.2 and present a Bayesian alternative for the estimation techniques discussed in Illustration 3.3. Programming code is in XXX and includes code for WinBUGS as well as the interface between R and WinBUGS, i.e. bugs from library BRugs, and its interaction with R library glmmBUGS).

[Reference to availability of programming code.]

Table 3 and Figure 7 summarize simulations from the posterior distribution of β_0 , β_1 and θ_0 in the varying intercepts Poisson GLMM specified in (89). Flat priors are used for β_0 and β_1 and a $UN(0, 100)$ prior for $\sqrt{\theta_0}$.

Figure 7: Posterior simulations for parameters used in (89) (from left to right: β_0 , β_1 and θ_0), workers' compensation insurance (frequencies) introduced in Illustration 1.2. Results are based on 2 chains, 50,000 simulations each, thinning factor of 5 and burn-in of 2,000 simulations.

	Mean	Median	SD		2.5%	5%	95%	97.5%
			Chain 1	Chain 2				
β_0	-3.563	-3.562	0.0851	0.0848	-3.732	-3.704	-3.423	-3.398
β_1	0.00978	0.00983	0.00511	0.00514	-0.000215	0.00138	0.0181	0.0198
θ_0	0.838	0.828	0.121	0.121	0.630	0.657	1.055	1.105

Table 3: Credibility intervals for parameters used in (89), workers' compensation insurance (frequencies) introduced in Illustration 1.2. Results are based on 2 chains, 50,000 simulations each, thinning factor of 5 and burn-in of 2,000 simulations.

Table 4 and Figure 8 summarize simulations from the posterior distribution of β_0 , β_1 and the variance components θ_0 , θ_1 and $\theta_{0,1}$ in model 90. Flat normal priors are used for β_0 and β_1 , together with a Wishart distribution for the inverse of $\mathbf{D}$.

With respect to predictive modeling, a Bayesian approach is most useful, since it provides the full predictive distribution of variables of interest (here: Count_{i7}). We illustrate this in Figure 9 for a selection of risk classes. Histograms are based on 50,000 simulations from the relevant predictive distribution (using model (90)). For each risk

Figure 8: Posterior simulations for parameters used in (90) (from left to right: β_0 , β_1 , θ_0 , θ_1 and $\theta_{0,1}$), workers' compensation insurance (frequencies) introduced in Illustration 1.2. Results are based on 2 chains, 50,000 simulations each, thinning factor of 5 and burn-in of 2,000 simulations.

	Mean	Median	SD		2.5%	5%	95%	97.5%
			Chain 1	Chain 2				
β_0	-3.581	-3.580	0.0849	0.0854	-3.750	-3.722	-3.442	-3.414
β_1	0.00675	0.00679	0.0176	0.0175	-0.0282	-0.0224	0.0354	0.0406
θ_0	0.838	0.827	0.119	0.118	0.635	0.662	1.047	1.101
θ_1	0.0244	0.024	0.00407	0.00410	0.0176	0.0185	0.0317	0.0335
$\theta_{0,1}$	0.000875	0.000903	0.0173	0.0171	-0.0332	-0.0271	0.0289	0.0348

Table 4: Credibility intervals for parameters used in (90), workers' compensation insurance (frequencies) introduced in Illustration 1.2. Results are based on 2 chains, 50,000 simulations each, thinning factor of 5 and burn-in of 2,000 simulations.

class the observed number of claims is indicated, as well as the point prediction obtained with the frequentist approach from Illustration 3.3 (using Laplace approximation from *glmer*).

Figure 9: Posterior predictive simulations for the number of claims in year 7 for a selection of risk classes. Simulations are based on (90), using 2 chains, 50,000 simulations each, thinning factor of 5 and burn-in of 2,000 simulations.

References

- M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions: with formulas, graphs and mathematical tables*. Dover, New York, 1972.
- K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58–76, 2007.
- K. Antonio and J. Beirlant. Issues in claims reserving and credibility: a semiparametric approach with mixed models. *Journal of Risk and Insurance*, 75(3):643–676, 2008.
- K. Antonio and E. Valdez. Statistical aspects of *a priori* and *a posteriori* risk classification in insurance. *Advances in Statistical Analysis*, 96(2):187–224, 2012.
- K. Antonio, E.W. Frees, and E. Valdez. A multilevel analysis of intercompany claim counts. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 40(1): 151–177, 2010.
- N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.
- N.E. Breslow and X. Lin. Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91, 1995.
- H. Bühlmann and A. Gisler. *A course in credibility theory and its applications*. Springer Verlag, Berlin, 2005.

- D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49:1–39, 1987.
- C. Czado. *Linear Mixed Models*. Lecture slides on GLM, TU Munchen, 2004.
- M. Davidian and R. J. Carroll. Variance function estimation. *Journal of the American Statistical Association*, 82:1079–1091, 1987.
- M. Denuit, X. Maréchal, S. Pitrebois, and J.-F. Walhin. *Actuarial Modelling Of Claim Counts: Risk Classification, Credibility and Bonus-Malus Scales*. Wiley, 2007.
- P. K. Dunn and G. K. Smyth. Series evaluation of tweedie exponential dispersion models densities. *Statistics and Computing*, 15:267–280, 2005.
- P. K. Dunn and G. K. Smyth. Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing*, 18:73–86, 2008.
- E.W. Frees. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press, Cambridge, 2004a.
- E.W. Frees. *Longitudinal and panel data. Analysis and applications in the social sciences*. Cambridge University Press, 2004b.
- E.W. Frees, V.R. Young, and Y. Luo. A longitudinal data analysis interpretation of credibility models. *Insurance: Mathematics and Economics*, 24(3):229–247, 1999.
- E.W. Frees, V.R. Young, and Y. Luo. Case studies using panel data models. *North American Actuarial Journal*, 5(4):24–42, 2001.
- A. Gelman and J. Hill. *Applied Regression and Multilevel (Hierarchical) Models*. Cambridge University Press, Cambridge, 2007.
- S. Haberman and A.E. Renshaw. Generalized linear models and actuarial science. *The Statistician*, 45(4):407–436, 1996.
- J.C. Hickman and L. Heacox. Credibility theory: the cornerstone of actuarial science. *North American Actuarial Journal*, 3(2):1–8, 1999.
- R.N. Kacker and D.A. Harville. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79:853–862, 1984.
- S. Klugman. *Bayesian statistics in actuarial science with emphasis on credibility*. Kluwer, Boston, 1992.
- N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.
- K.Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.
- X. Lin and N.E. Breslow. Analysis of correlated binomial data in logistic-normal models. *Journal of Statistical Computation and Simulation*, 55:133–146, 1996.

- Q. Liu and D.A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):624–629, 1994.
- U. Makov, A.F.M. Smith, and Y.H. Liu. Bayesian methods in actuarial science. *The Statistician*, 45(4):503–515, 1996.
- C.E. McCulloch and S.R. Searle. *Generalized, Linear and Mixed Models*. Wiley Series in Probability and Statistics, Wiley, New York, 2001.
- G. Molenberghs and G. Verbeke. *Models for discrete longitudinal data*. Springer Series In Statistics, Springer, New York, 2005.
- J.A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74: 221–232, 1987.
- J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.
- G. W. Peters, P. V. Shevchenko, and M. V. Wüthrich. Model uncertainty in claims reserving within tweedie’s compound poisson models. *ASTIN Bulletin*, 39:1–33, 2009.
- J.C. Pinheiro and D.M. Bates. *Mixed Effects Models in S and S-Plus*. Springer, 2000.
- O. Purcaru, M. Guillén, and M. Denuit. Linear credibility models based on time series for claim counts. *Belgian Actuarial Bulletin*, 4(1):62–74, 2004.
- G.K. Robinson. That blup is a good thing: the estimation of random effects. *Statistical Science*, 6:15–51, 1991.
- D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. Cambridge University Press, Cambridge, 2003.
- D. Scollnik. An introduction to Markov Chain Monte Carlo methods and their actuarial applications. *Proceedings of the Casualty Actuarial Society Forum*, LXXXIII:114–165, 1996.
- S.R. Searle, G. Casella, and C.E. McCulloch. *Variance components*. Wiley, 2008.
- L. Tierny and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.
- F. Tuerlinckx, F. Rijmen, G. Verbeke, and P. De Boeck. Statistical inference in generalized linear mixed models: a review. *British Journal of Mathematical and Statistical Psychology*, 59:225–255, 2006.
- G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer Series In Statistics, New York, 2000.
- R. Wolfinger and M. O’Connell. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243, 1993.
- S. Wood. *Generalized Additive Models: An introduction with R*. Chapman & Hall, CRC Texts in Statistical Science, 2006.

- K. C. H. Yip and K. K. W. Yau. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36:153–163, 2005.
- Y. Zhang. Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models. *Statistics and Computing*, 2012. forthcoming.
- Y. Zhang and V. Dukic. Predicting multivariate insurance loss payments under the bayesian copula framework. *The Journal of Risk and Insurance*, 2012. in press, DOI: 10.1111/j.1539-6975.2012.01480.x.
- Y. Zhang, V. Dukic, and J. Guszcz. A bayesian nonlinear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society, Series A*, 175:637–656, 2012.
- Y. Zhao, J. Staudenmayer, B.A. Coull, and M.P. Wand. General design bayesian generalized linear mixed models. *Statistical Science*, 21:35–51, 2006.

FACULTY OF ECONOMICS AND BUSINESS

Naamsestraat 69 bus 3500

3000 LEUVEN, BELGIË

tel. + 32 16 32 66 12

fax + 32 16 32 67 91

info@econ.kuleuven.be

www.econ.kuleuven.be