KU LEUVEN
FACULTY OF ECONOMICS AND BUSINESS

Linear mixed models for predictive modelling in actuarial science

Katrien Antonio, Yanwei Zhang

AFI_1385

Linear mixed models for predictive modelling in actuarial science

Katrien Antonio [*] Yanwei Zhang [†]

November 24, 2013

Chapter preview. We give a general discussion of linear mixed models and continue with illustrating specific actuarial applications of this type of models. Technical details on linear mixed models follow: model assumptions, specifications, estimation techniques and methods of inference. We include three worked out examples with the `R` `lme4` package and use `ggplot2` for the graphs. Full code is available from the book project's web page.

1 Mixed models in actuarial science

1.1 What?

A first example of a linear mixed model. As explained in Chapter XXX , a panel data set follows a group of subjects (e.g. policyholders in an insurance portfolio) over time. We therefore denote variables (e.g. y_{it}, $\boldsymbol{x}_{it}$) in a panel data set with double subscripts, indicating the subject (say i) and the time period (say t). As motivated in Section 1.2 of Chapter XXX, the analysis of panel data has several advantages. Panel data allow to study the effect of certain covariates on the response of interest (as in usual regression models for cross–sectional data), while accounting appropriately for the dynamics in these relations. For actuarial ratemaking the availability of panel data is of particular interest in light of *a posteriori* rating. An *a posteriori* tariff predicts the current year loss for a particular policyholder, using (among others) the dependence between the current year's loss and losses reported by this policyholder in previous years. Credibility theory, being a cornerstone of actuarial mathematics, is an example of such an *a posteriori* rating system. Section 2 in Chapter XXX presents a sequence of models suitable for the analysis of panel data in the context of linear models. Recall in particular the well–known linear regression model with common intercept (or: cross–sectional model) (see 'Linear Model 1' in Chapter XXX, Section 2)

[Reference to Chapter on longitudinal data.]

$$Ey_{it} = \alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta}. \tag{1}$$

[*] University of Amsterdam and KU Leuven (Belgium), email: k.antonio@uva.nl

[†] University of Southern California, email: actuary_zhang@hotmail.com

This model *completely pools* the data, ignores the panel structure and produces identical estimates for all subjects i (for a given x_{it}). The linear fixed effects model ('Linear Model 2' in Chapter XXX, Section 2) specifies

$$Ey_{it} = \alpha_i + \boldsymbol{x}_{it}^{'}\boldsymbol{\beta}, \tag{2}$$

where each subject i has its own unknown - but *fixed* - intercept α_i. Hence, the name **fixed effects** model. Independence among all observations is assumed, and $\text{Var}(y_{it}) = \sigma^2$. This regression model *does not pool* information and estimates each α_i separately using least squares or maximum likelihood. This approach often results in overfitting and unreasonable $\hat{\alpha}_i^{'}s$ (see Gelman (2006)). The linear random effects model (see 'Linear Model 3' in Chapter XXX) is an alternative approach, balancing between *no pooling* and *complete pooling* of data. It allows for *random* intercepts, with model equation

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \epsilon_{it}, \tag{3}$$

where $\epsilon_{it} \sim (0, \sigma_\epsilon^2)$ [1]. The subject specific intercept α_i is now a random variable with zero mean and variance σ_α^2. Hence the name **random effects** model. Moreover, the model in (3) is a first example of a **linear mixed model** ([LMM]), with a combination (*'mix'*) of *fixed* and *random* effects in the linear predictor. The errors ϵ_{it} with variance σ_ϵ^2 structure variability within subject i, whereas the random intercepts with variance σ_α^2 represent variation between subjects. Compared with the *no pooling* and *complete pooling* examples, the linear mixed model has many interesting features, as is explained below.

[Reference to Chapter on longitudinal and panel data]

Mixed or multilevel models for clustered data. Panel data is a first example of so–called clustered data. As mentioned in Section 4 in Chapter XXX , predictive modeling in actuarial science (and in many other statistical disciplines) will confront analysts with data structures going beyond the cross–sectional as well as panel data design. Section 3 in this chapter includes multiple motivating examples. Mixed (or: multilevel) models are statistical models suitable for the analysis of data structured in nested (i.e. *hierarchical*) or non–nested (i.e. cross–classified, *next to* each other instead of hierachically nested) **clusters or levels**. In this chapter we explain the use of **linear mixed models** for multilevel data. A discussion of **non–linear** mixed models follows in Chapter XXX . Chapter XXX (on longitudinal and panel data), XXX (on credibility), and XXX (on spatial statistics) in this book include additional examples of clustered data and their analysis with mixed models.

[Reference to Chapter on non-linear mixed models.]

[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 $j = 1, \ldots, n_i$) insured under fleets i (with $i = 1, \ldots, m$). Let y_{ij} be the loss observed

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

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)

$$y_{ij} \quad = \quad \beta_i + \beta_{1,0} + x_{1,ij}\beta_{1,1} + \epsilon_{1,ij}, \qquad j = 1, \ldots, n_i, \tag{4}$$

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

$$\beta_i \quad = \quad \epsilon_{2,i}, \qquad i = 1, \ldots, m, \tag{5}$$

or, when fleet–specific information is available,

$$\beta_i \quad = \quad x_{2,i}\beta_2 + \epsilon_{2,i}, \qquad i = 1, \ldots, m. \tag{6}$$

Here $\epsilon_{2,i} \sim (0, \sigma_2^2)$ 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. Written as a **single model equation**, the combination of (4) and, for example, (5), is:

$$y_{ij} \quad = \quad \beta_{1,0} + \epsilon_{2,i} + x_{1,ij}\beta_{1,1} + \epsilon_{1,ij}. \tag{7}$$

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 (i.e. the variance component σ_2^2) in their distribution. 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 (7) is – again – an 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

$$y_{ij} \quad = \quad \beta_{i,0} + x_{1,ij}\beta_{i,1} + \beta_{1,0} + x_{1,ij}\beta_{1,1} + \epsilon_{1,ij}, \qquad i = 1, \ldots, m, \; j = 1, \ldots, n_i, \tag{8}$$

with

$$\begin{aligned} \beta_{i,0} \quad &= \quad \epsilon_{2,i,0}, \\ \beta_{i,1} \quad &= \quad \epsilon_{2,i,1}. \end{aligned} \tag{9}$$

Written as a single model equation, this multilevel model becomes

$$y_{ij} \;=\; \beta_{1,0} + \epsilon_{2,i,0} + x_{1,ij}\beta_{1,1} + x_{1,ij}\epsilon_{2,i,1} + \epsilon_{1,ij}. \tag{10}$$

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 slopes $\epsilon_{2,i,1}$.

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, policyholder data over time, policies within counties, ...), statistical methodology should reflect the structure in the data and use it as relevant information when building statistical models. Using traditional (say linear or generalized linear models, as in Chapter XXX and XXX) regression techniques, the clustering in groups is either ignored (***'complete pooling'***) or groups are analyzed separately (***'no pooling'***) resulting in overfitting because even small clusters will get their own regression model. The multilevel model enhances both extremes, e.g. in the varying intercepts model from (7) 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 $\epsilon_{2,i}$ (with variance σ_2^2) 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 (6)). 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 by 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 (or: cluster) effects, predictor information at different levels, and so on. For instance, Gelman and Hill (2007) denote the varying coefficients and varying slopes models in (4)+(6) and (10), respectively, in a more intuitive way:

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, N \\ \alpha_j &= a + b u_j + \eta_j, \qquad j = 1, \ldots, m, \end{aligned} \tag{11}$$

and

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i, \qquad i = 1, \ldots, N \\ \alpha_j &= a_0 + b_0 u_j + \eta_{j1}, \qquad j = 1, \ldots, m \\ \beta_j &= \eta_{j2}. \end{aligned} \tag{12}$$

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 (4)) and u_j refers to covariate information available at fleet–level (i.e. level 2 in (6)).

The notation used from Section 2 on is motivated by generality, and inspired by Frees (2004a). 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. Section 2 explains 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. We will deploy some of these examples within the framework of linear mixed models, while others are more appropriate for analysis with generalized linear mixed models (see Chapter XXX).

[Reference to Chapter on non–linear mixed models.]

Illustration 1 (Credibility models). *Credibility theory is an a posteriori ratemaking technique. Credibility models are designed for the prediction of an insured's risk premium, by weighting the insured's own loss experience and the experience in the overall portfolio. An extensive discussion of credibility models is available in Chapter XXX in this book. Credibility models have a natural and explicit interpretation as special examples of mixed models. Frees et al. (1999) demonstrate this connection, by reinterpreting credibility models using mixed model parlance. This mapping highly increases the accessibility and usefulness of such models. Indeed, the complete machinery (including computational methods and soft-*

[Reference to Chapter on credibility.]

ware) of mixed models becomes available for the analysis of these actuarial models. The famous Hachemeister data set (see Hachemeister (1975)) has often been used in credibility literature. This data set considers 12 periods, from the third quarter of 1970 to the second quarter of 1973, of bodily injury losses covered by a private passenger auto insurance. For 5 states the total loss and corresponding number of claims are registered. Figure 1 shows a trellis plot (see Chapter XXX) of the average loss per claim (in black), followed over time, per state. The plot also shows a linear regression line (in blue) and corresponding confidence intervals (in grey). In Section 3 we use linear mixed models to analyze this data set and predict the next year's average claim per state. Further analysis – with focus on credibility theory – follows in Chapter XXX.

[Reference to Chapter on longitudinal and panel data.]

[Reference to Chapter on credibility.]

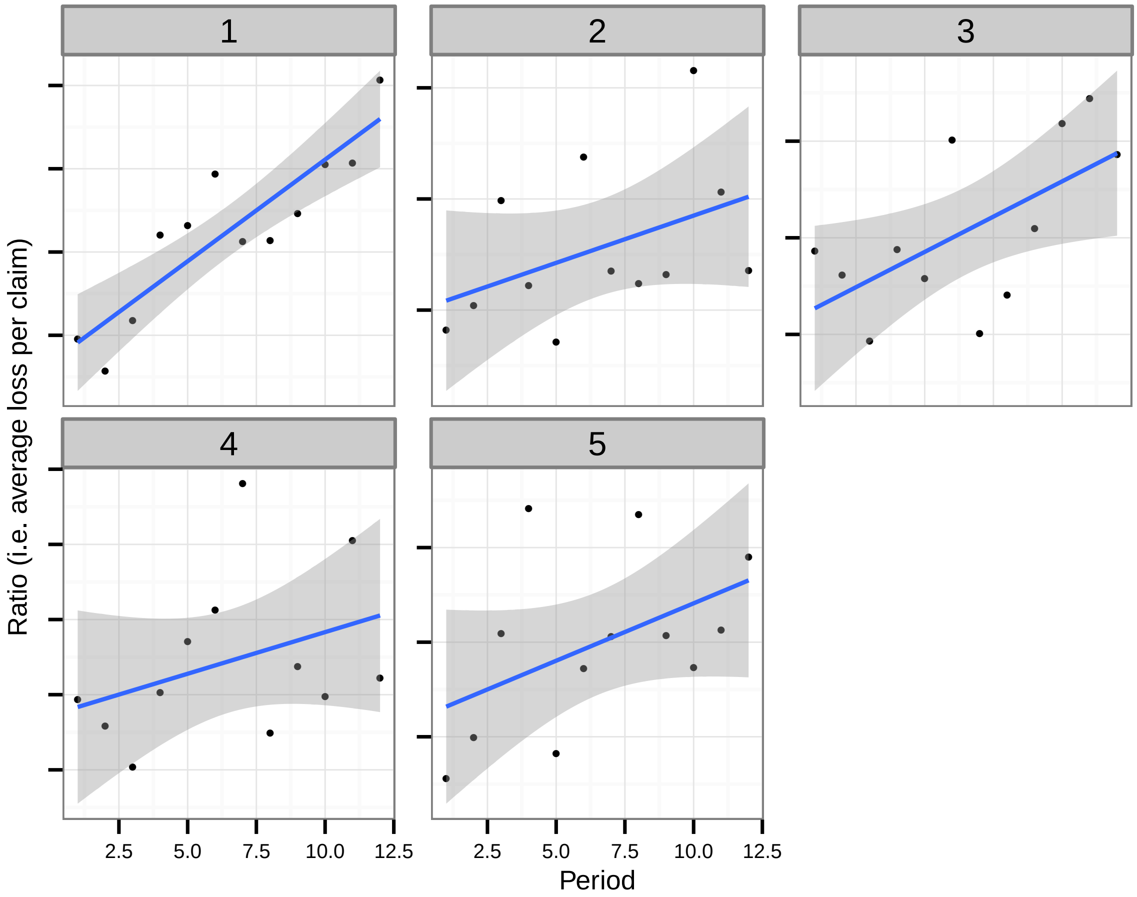

Figure 1: *Trellis plot of average losses per period (in black) and a linear regression line (in blue) with corresponding confidence intervals (in grey); each panel represents one state: Hachemeister data.*

Illustration 2 (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 of interest is the* Loss *paid out (on a yearly basis) per risk* *class. Possible explanatory variables are* Year *and* Payroll*. Frees et al. (2001) and Antonio and Beirlant (2007) present mixed models for the pure premium,* PP=Loss/Payroll*. For a random subsample of 10 risk classes, Figure 2 shows the time series plot of* Loss *(left) and corresponding* Payroll *(right).*

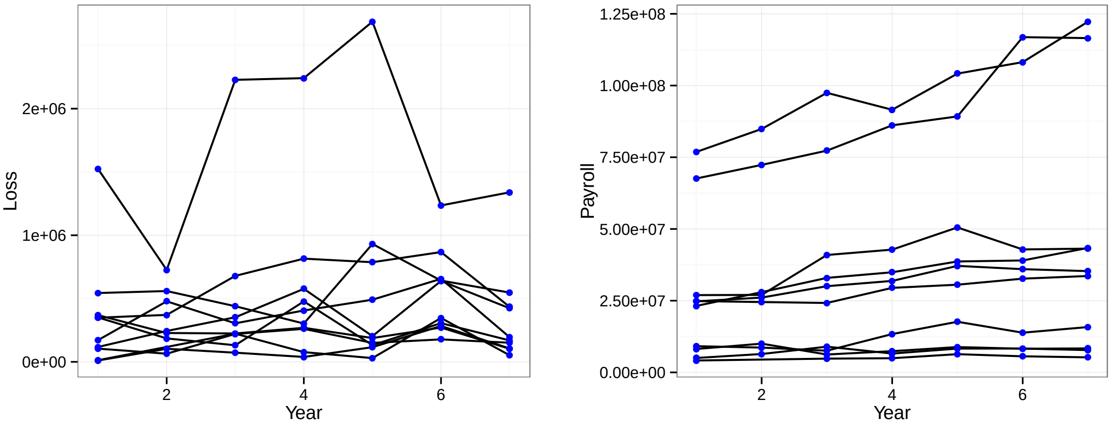

Figure 2: *Time series plot of losses (left) and payroll (right) for a random sample of 10 risk classes: workers' compensation data (losses).*

Illustration 3 (Workers' Compensation Insurance: frequencies)**.** *The data are from Klugman (1992) (see Scollnik (1996), Makov et al. (1996) and Antonio and Beirlant (2007) for further discussion). Frequency counts in workers' compensation insurance are observed on a yearly basis for 133 occupation classes followed during 7 years.* Count *is the response variable of interest. Possible explanatory variables are* Year *and* Payroll*, a measure of exposure denoting scaled payroll totals adjusted for inflation. Figure 3 shows exploratory plots for a random subsample of 10 occupation classes. Statistical modeling should take into account the dependence between observations on the same occupation class 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 the observed claims history of this particular risk class and the whole portfolio?'. Since the response variable in this example is claim frequency, we will analyze this data set within the context of Generalized Linear Mixed instead of Linear Mixed Models (in Chapter XXX).*

[Reference to Chapter on non–linear mixed models.]

Illustration 4 (Hierarchical data structures)**.** *With panel data a group of subjects is followed over time, as in Illustrations 2 and 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 employees 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*

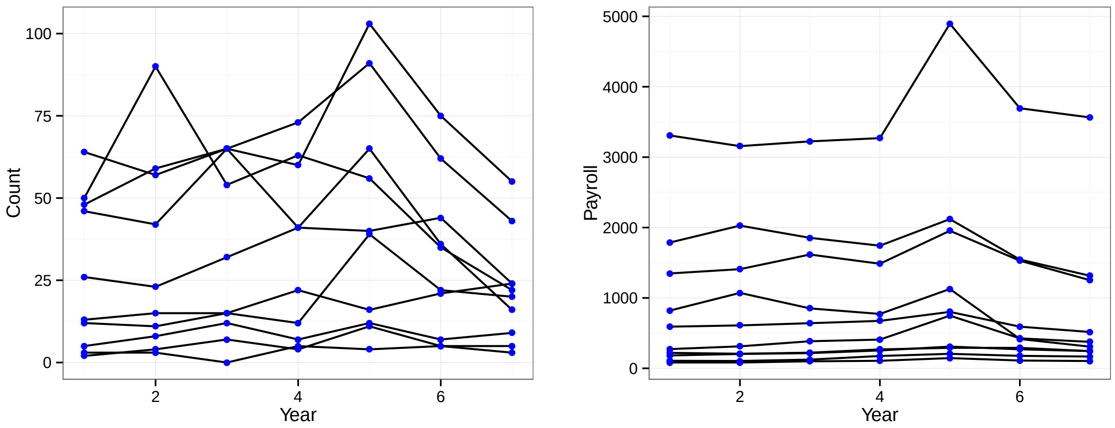

Figure 3: *Time series plot of counts (left) and payroll (right) for a random sample of 10 risk classes: workers' compensation data (counts).*

*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 5 (Non–nested or cross–classified data structures)**.** *Data may also be structured in levels which are not nested or hierarchically structured, but instead act next to each other. An example is the data set from Dannenburg et al. (1996) on private loans from a credit insurer. The data are payments of the credit insurer to several banks for covering losses caused by clients who were no longer able to pay off their loans. These payments are categorized by civil* status *of the debtors and their work* experience*. The civil status is single (1), divorced (2) or other (3), and the work experience is less than two* *years (< 2, category 1), from 2 up to 10 years (≥ 2 and < 10, category 2) and more than then years (≥ 10, category 3). Table 1 shows the number of clients and the average loss paid per risk class.*

	experience		
status	1	2	3
1	40	43	41
2	54	53	48
3	39	39	44

	experience		
status	1	2	3
1	180.39	246.71	261.58
2	172.05	232.67	253.22
3	212.30	269.56	366.61

Table 1: *Number of payments (left) and average loss per combination of* status *and* experience *risk class: credit insurance data.*

Boxplots of the observed payments per risk class are in Figure 4. Using linear mixed models we estimate the expected loss per risk category, and compare our results with the credibility premiums derived by Dannenburg et al. (1996).

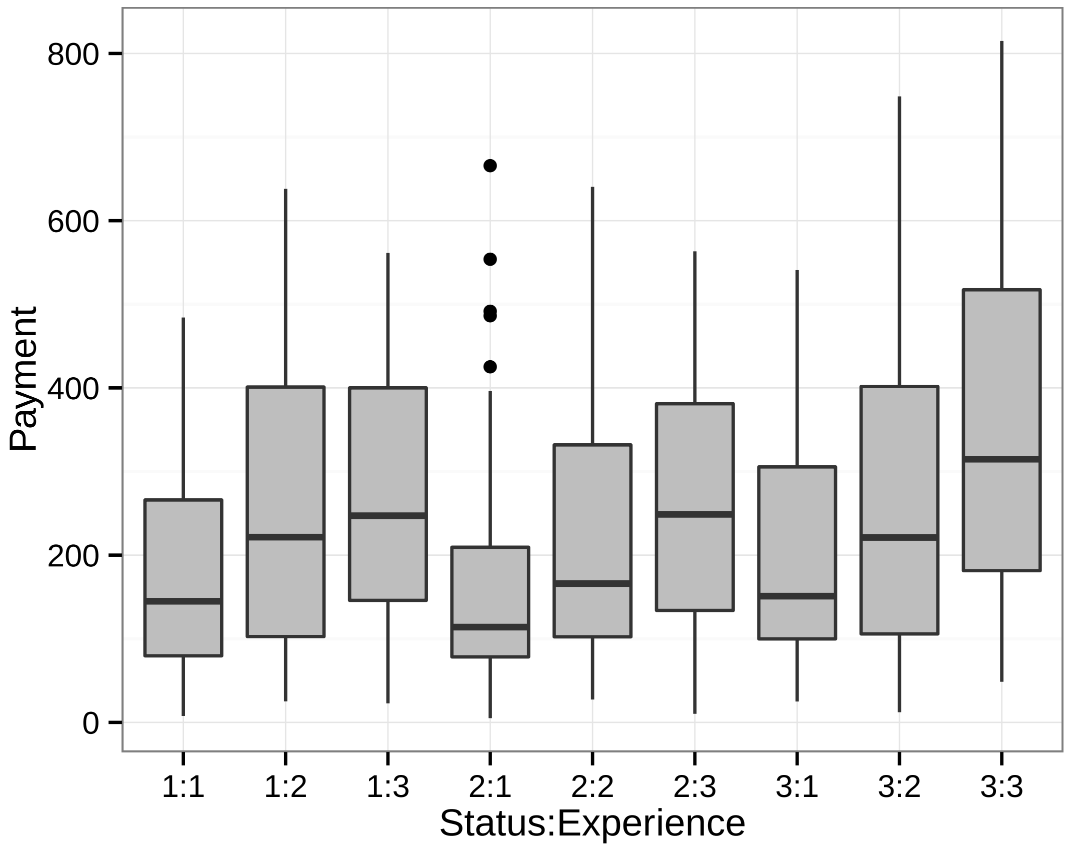

Figure 4: *Boxplots of payments versus combination of status and experience: credit insurance data.*

Illustration 6 (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 Commissioners* [2]. *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 5. Predicting the development of these losses beyond the range of the available data, is the major challenge in loss reserving. Figure 5 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.*

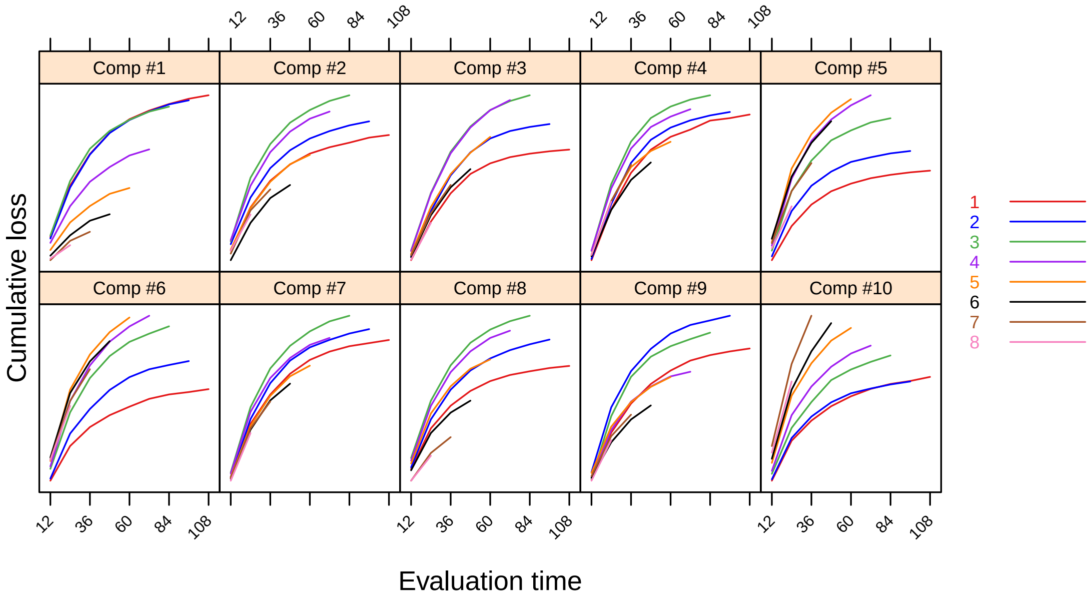

Figure 5: *Observed growth of cumulative losses for the 10 companies in study. The colored lines represent accident years.*

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 (2004a).

[2] NAIC is a consortium of state–level insurance regulators in the United States.

2.1 Model assumptions and notation

The basic linear model specifies $E[\boldsymbol{Y}] = \boldsymbol{X\beta}$ with $\boldsymbol{Y}$ the response vector, $\boldsymbol{\beta}$ the vector of regression parameters and $\boldsymbol{X}$ the model design matrix. 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 $\boldsymbol{X\beta}$, but add $\boldsymbol{Zu}$ to it, where $\boldsymbol{Z}$ is a model matrix, corresponding with a vector of random effects $\boldsymbol{u}$. A distribution is specified for this random effects vector $\boldsymbol{u}$ with mean zero and covariance matrix $\boldsymbol{D}$. 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]

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon} \\ \boldsymbol{u} &\sim (\boldsymbol{0}, \boldsymbol{D}) \\ \boldsymbol{\epsilon} &\sim (\boldsymbol{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{13}$$

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

$$\begin{aligned} E[\boldsymbol{y}|\boldsymbol{u}] &= \boldsymbol{X\beta} + \boldsymbol{Zu}, \\ \text{Var}[\boldsymbol{y}|\boldsymbol{u}] &= \boldsymbol{\Sigma}. \end{aligned} \tag{14}$$

The combined, unconditional or ***marginal*** model states

$$\boldsymbol{y} \sim (\boldsymbol{X\beta}, \boldsymbol{V} := \boldsymbol{ZDZ}' + \Sigma), \tag{15}$$

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

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

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} \end{pmatrix} \right). \tag{16}$$

With these distributional assumptions the hierarchical LMM becomes

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{u} &\sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \boldsymbol{\Sigma}) \\ \boldsymbol{u} &\sim N(\boldsymbol{0}, \boldsymbol{D}). \end{aligned} \tag{17}$$

This implies the marginal model $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$, 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 $\boldsymbol{u}$ the specification in (13) and (17) should be used.

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

Illustrations 7 and 8 below focus on particular examples of 2 and 3 level data and explain in detail the structure of vectors and matrices in (13) and (15).

Illustration 7 (A 2–level model for longitudinal data.)**.** *Y_{ij} represents the jth measurement on a subject i (with $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$). m is the number of subjects under consideration and n_i the number of observations registered on subject i. $\boldsymbol{x}_{ij}$ ($p \times 1$) is a column vector with fixed effects' covariate information from observation j on subject i. Correspondingly, $\boldsymbol{z}_{ij}$ ($q \times 1$) is a column vector with covariate information corresponding with random effects. $\boldsymbol{\beta}$ ($p \times 1$) is a column vector with fixed effects parameters and $\boldsymbol{u}_i$ ($q \times 1$) is a column vector with random effects regression parameters. These are subject–specific and allow to model heterogeneity between subjects. The combined model is*

$$y_{ij} \;=\; \underbrace{\boldsymbol{x}^{'}_{ij}\boldsymbol{\beta}}_{\textit{fixed}} + \underbrace{\boldsymbol{z}^{'}_{ij}\boldsymbol{u}_i}_{\textit{random}} + \underbrace{\epsilon_{ij}}_{\textit{random}} \;. \tag{18}$$

The distributional assumptions for the random parts in (18) *are*

$$\begin{aligned} \boldsymbol{u}_i &\sim (\mathbf{0}, \boldsymbol{G}) \qquad \boldsymbol{G} \in \mathbb{R}^{q \times q} \\ \boldsymbol{\epsilon}_i &\sim (\mathbf{0}, \boldsymbol{\Sigma}_i) \qquad \boldsymbol{\Sigma}_i \in \mathbb{R}^{n_i \times n_i}. \end{aligned} \tag{19}$$

The covariance matrix $\boldsymbol{G}$ is left unspecified, i.e. no particular structure is implied. Various structures are available for $\boldsymbol{\Sigma}_i$. Very often just a simple diagonal matrix is used: $\boldsymbol{\Sigma}_i := \sigma^2 I_{n_i}$. 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 $\boldsymbol{\Sigma}_i$ as non–diagonal (e.g. unstructured, Toeplitz or autoregressive structure, see Verbeke and Molenberghs (2000) for more discussion). $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_m$ are independent. Typically, normality is assumed for both vectors, as in (17)*. In vector notation we specify*

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, m, \\ \boldsymbol{u}_i &\sim (\mathbf{0}, \boldsymbol{G}) \\ \boldsymbol{\epsilon}_i &\sim (\mathbf{0}, \boldsymbol{\Sigma}_i), \end{aligned} \tag{20}$$

where

$$\boldsymbol{X}_i := \begin{pmatrix} \boldsymbol{x}^{'}_{i1} \\ \vdots \\ \boldsymbol{x}^{'}_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times p}, \quad \boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{z}^{'}_{i1} \\ \vdots \\ \boldsymbol{z}^{'}_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times q}, \quad \boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times 1}. \tag{21}$$

Combining all subjects or clusters $i = 1, \ldots, m$, (13) is the matrix formulation of this LMM for longitudinal data (with $N = \sum_{i=1}^{m} n_i$ the total number of observations)

$$\begin{aligned}
\boldsymbol{y} &= \begin{pmatrix} \boldsymbol{y}_1 \\ \ldots \\ \boldsymbol{y}_m \end{pmatrix} \in \mathbb{R}^{N\times 1}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix} \in \mathbb{R}^{N\times p}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix} \in \mathbb{R}^{N\times 1}, \\
\boldsymbol{Z} &= \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0}_{n_1\times q} & \ldots & \boldsymbol{0}_{n_1\times q} \\ \boldsymbol{0}_{n_2\times q} & \boldsymbol{Z}_2 & & \\ \vdots & & \ddots & \\ \boldsymbol{0}_{n_m\times q} & & & \boldsymbol{Z}_m \end{pmatrix} \in \mathbb{R}^{N\times (m\cdot q)}, \quad \boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{pmatrix} \in \mathbb{R}^{(m\cdot q)\times 1}. \quad (22)
\end{aligned}$$

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

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{G} & & \\ & \ddots & \\ & & \boldsymbol{G} \end{pmatrix} \in \mathbb{R}^{m\cdot q\times m\cdot q}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_m \end{pmatrix} \in \mathbb{R}^{N\times N}. \quad (23)$$

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

$$\begin{aligned}
\boldsymbol{V} &= \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{\Sigma} \\
&= \begin{pmatrix} \boldsymbol{Z}_1\boldsymbol{G}\boldsymbol{Z}_1' + \boldsymbol{\Sigma}_1 & \ldots & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{Z}_m\boldsymbol{G}\boldsymbol{Z}_m' + \boldsymbol{\Sigma}_m \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{V}_1 & & \\ & \ddots & \\ & & \boldsymbol{V}_m \end{pmatrix}, \quad (24)
\end{aligned}$$

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

Illustration 8 (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:*

$$y_{ijk} = \boldsymbol{z}_{1,ijk}'\boldsymbol{\beta}_{ij} + \boldsymbol{x}_{1,ijk}'\boldsymbol{\beta}_1 + \epsilon_{1,ijk}. \quad (25)$$

Hereby, predictors $\boldsymbol{z}_{1,ijk}$ and $\boldsymbol{x}_{1,ijk}$ 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:

$$\boldsymbol{\beta}_{ij} = \boldsymbol{Z}_{2,ij}\boldsymbol{\gamma}_i + \boldsymbol{X}_{2,ij}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_{2,ij}. \quad (26)$$

$\boldsymbol{X}_{2,ij}$ and $\boldsymbol{Z}_{2,ij}$ may depend on company or fleet, but not on the insured vehicle. The regression parameters in $\boldsymbol{\gamma}_i$ are company–specific and modeled in (27)*:*

$$\boldsymbol{\gamma}_i \;=\; \boldsymbol{X}_{3i}\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}_{3i}, \tag{27}$$

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

$$\begin{aligned} Y_{ijk} &= z^{'}_{1,ijk}(\boldsymbol{Z}_{2,ij}(\boldsymbol{X}_{3,i}\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}_{3i}) + \boldsymbol{X}_{2,ij}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_{2,ij}) + \boldsymbol{x}^{'}_{1,ijk}\boldsymbol{\beta}_1 + \epsilon_{1,ijk} \\ &= \boldsymbol{x}^{'}_{ijk}\boldsymbol{\beta} + \boldsymbol{z}^{'}_{ijk}\boldsymbol{u}_{ij} + \epsilon_{1,ijk}, \end{aligned} \tag{28}$$

where $\boldsymbol{x}^{'}_{ijk} = (\boldsymbol{x}^{'}_{1,ijk} \quad \boldsymbol{z}^{'}_{1,ijk}\boldsymbol{X}_{2,ij} \quad \boldsymbol{z}^{'}_{1,ijk}\boldsymbol{Z}_{2,ij}\boldsymbol{X}_{3i})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^{'}_1 \quad \boldsymbol{\beta}^{'}_2 \quad \boldsymbol{\beta}^{'}_3)^{'}$, $\boldsymbol{z}^{'}_{i,j,k} = (\boldsymbol{z}^{'}_{1,i,j,k} \quad \boldsymbol{z}^{'}_{1,i,j,k}\boldsymbol{Z}_{2,ij})$ and $\boldsymbol{u}_{ij} = (\boldsymbol{\epsilon}^{'}_{2,ij} \quad \boldsymbol{\epsilon}^{'}_{3,i})^{'}$. 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 (2004a) (Appendix 5A).

2.2 The structure of random effects

Since the random effects $\boldsymbol{u}$ often correspond to factor predictors, the design matrix $\boldsymbol{Z}$ is often highly sparse, with a high proportion of elements to be exactly zero. Moreover, the covariance matrix $\boldsymbol{D}$ is highly structured and depends on some parameter vector $\boldsymbol{\theta}$ that is to be estimated.

- **Single random effect per level**. This is the simplest yet most common case where the random effect corresponds to a certain level of a single grouping factor. For example, we may have the state indicator in the model and each state has its own intercept, i.e. `y ~ (1|state)` (in `R` parlance). We illustrate this structure in Section 3 with the workers' compensation losses data.

- **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 also its own slope with respect to some predictor, i.e., `y ~ (1 + time|state)`. In general, the multiple random effects are correlated, and so the matrix $\boldsymbol{D}$ is not diagonal. We illustrate this structure in Section 3 with the workers' compensation losses data.

- **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, i.e., `y ~ (1|county/state)`. 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, i.e., `y ~ (1|state) + (1|make)`. The credit insurance example in Section 3 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 (4) and (5), $\beta_{1,0}$ and $\beta_{1,1}$ are regression parameters corresponding with fixed effects, σ_1^2 and σ_2^2 are variance components and $\epsilon_{2,i}$ $(i = 1, \ldots, m)$ 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 $\boldsymbol{\beta}$ and a predictor for the random effects in $\boldsymbol{u}$, under the assumption of known covariance parameters in $\boldsymbol{V}$ (see (15)).

Estimating $\boldsymbol{\beta}$. The Generalized Least Squares ([GLS]) estimator – which coincides with the maximum likelihood estimator ([MLE]) under normality (as in (17)) – of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{29}$$

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

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

$$\text{BLP}[\boldsymbol{u}] \;=\; \hat{\boldsymbol{u}} = E[\boldsymbol{u}] + \boldsymbol{C}\boldsymbol{V}^{-1}(\boldsymbol{y} - E[\boldsymbol{y}]), \tag{30}$$

where $\boldsymbol{V} = \text{Var}(\boldsymbol{y})$ and $\boldsymbol{C} = \text{Cov}(\boldsymbol{u}, \boldsymbol{y}')$. BP($\boldsymbol{u}$) and BLP($\boldsymbol{u}$) are unbiased, in the sense that their expected value equals $E[\boldsymbol{u}]$. Normality is not required in BP or BLP, but with $(\boldsymbol{y}\;\;\boldsymbol{u})$ 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 (17) the predictor of $\boldsymbol{u}$ is usually called the Best Linear Unbiased Predictor ([BLUP]). Robinson (1991) describes several ways to derive this BLUP. For instance, under normality assumptions:

$$\begin{aligned} Cov(\boldsymbol{y}, \boldsymbol{u}^{'}) &= Cov(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}, \boldsymbol{u}^{'}) \\ &= Cov(\boldsymbol{X\beta}, \boldsymbol{u}^{'}) + \boldsymbol{Z}\mathrm{Var}(\boldsymbol{u}, \boldsymbol{u}^{'}) + \mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{u}^{'}) \\ &= \boldsymbol{ZD}, \end{aligned}$$

which leads to the multivariate normal distribution

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{X\beta} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V} & \boldsymbol{ZD} \\ \boldsymbol{DZ}^{'} & \boldsymbol{D} \end{pmatrix} \right). \tag{31}$$

Using either properties of this distribution [4] or the result in (30) the BLUP of $\boldsymbol{u}$ follows:

$$\mathrm{BLUP}(\boldsymbol{u}) := \hat{\boldsymbol{u}} = \boldsymbol{DZ}^{'}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{32}$$

Of course, (32) relies on the (unknown) vector of fixed effects $\boldsymbol{\beta}$, as well as on unknown covariance parameters in $\boldsymbol{V}$. 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 $(\boldsymbol{\beta}, \boldsymbol{u})$ and from the estimation of $\boldsymbol{V}$. Histograms and scatter plots of components of $\hat{\boldsymbol{u}}$ 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 $(\boldsymbol{y}^{'}, \boldsymbol{u}^{'})^{'}$ (see assumptions (17)) with respect to $(\boldsymbol{\beta}, \boldsymbol{u})$ leads to Henderson's mixed model equations:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}) &= f(\boldsymbol{y}|\boldsymbol{u}) \cdot f(\boldsymbol{u}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^{'}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) \right) \cdot \exp(-\frac{1}{2}\boldsymbol{u}^{'}\boldsymbol{D}^{-1}\boldsymbol{u}). \end{aligned} \tag{33}$$

It is therefore enough to minimize

$$Q(\boldsymbol{\beta}, \boldsymbol{u}) := (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^{'}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \boldsymbol{u}^{'}\boldsymbol{Du}, \tag{34}$$

which corresponds to solving the set of equations

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}}Q(\boldsymbol{\beta}, \boldsymbol{u}) &= \boldsymbol{0} \text{ and } \frac{\partial}{\partial \boldsymbol{u}}Q(\boldsymbol{\beta}, \boldsymbol{u}) = \boldsymbol{0} \\ &\Leftrightarrow \begin{pmatrix} \boldsymbol{X}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{X}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{Z}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \end{pmatrix}. \end{aligned} \tag{35}$$

[4] Namely: with $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu_Y} \\ \boldsymbol{\mu_Z} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_Y} & \boldsymbol{\Sigma_{YZ}} \\ \boldsymbol{\Sigma_{ZY}} & \boldsymbol{\Sigma_Z} \end{pmatrix} \right)$ we know $\boldsymbol{Z}|\boldsymbol{Y} \sim N(\boldsymbol{\mu_{Z|Y}}, \boldsymbol{\Sigma_{Z|Y}})$ where $\boldsymbol{\mu_{Z|Y}} = \boldsymbol{\mu_Z} + \boldsymbol{\Sigma_{ZY}}\boldsymbol{\Sigma_Y}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_Y})$ and $\boldsymbol{\Sigma_{Z|Y}} = \boldsymbol{\Sigma_Z} - \boldsymbol{\Sigma_{ZY}}\boldsymbol{\Sigma_Y}^{-1}\boldsymbol{\Sigma_{YZ}}$.

(29) and (32) solve this system of equations.

More on prediction. With $\hat{\boldsymbol{\beta}}$ from (29) and $\hat{\boldsymbol{u}}$ from (32), the profile of cluster i is predicted by

$$\begin{aligned} \hat{\boldsymbol{y}}_i &:= \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i \\ &= \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{'}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\hat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i} - \boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1})\boldsymbol{Y}_i, \end{aligned} \tag{36}$$

using $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{'} + \boldsymbol{\Sigma}_i$ and n_i the cluster size. $\hat{\boldsymbol{y}}_i$ is a weighted mean of the global profile $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ and the data observed on cluster i, $\boldsymbol{y}_i$. $\hat{\boldsymbol{y}}_i$ is a so–called *shrinkage estimator*. Actuaries will recognize a credibility type formula in (36).

The prediction of a future observation is discussed in detail in Frees (2004b) (Section 4.4). The case of non–diagonal residual covariance matrices $\boldsymbol{\Sigma}_i$ requires special attention. For instance, with panel data the BLUP for y_{i,T_i+1} is $\boldsymbol{x}_{i,T_i+1}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i,T_i+1}^{'}\hat{\boldsymbol{u}}_i +$ BLUP(ϵ_{i,T_i+1}). From (30) we understand that the last term in this expression is zero when $Cov(\epsilon_{i,T_i+1}, \boldsymbol{\epsilon}_i) = \boldsymbol{0}$. 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.]

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

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \log\{L(\boldsymbol{\beta}, \boldsymbol{\theta})\} \\ &= -\frac{1}{2}\Big(\ln|\boldsymbol{V}(\boldsymbol{\theta})| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'}\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big) + c. \end{aligned} \tag{37}$$

Maximizing (37) with respect to $\boldsymbol{\beta}$ and with $\boldsymbol{\theta}$ fixed, we get

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}^{'}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{y}. \tag{38}$$

We obtain the so–called *profile log–likelihood* by replacing $\boldsymbol{\beta}$ in (37) with $\hat{\boldsymbol{\beta}}$ from (38)

$$\begin{aligned} \ell_p(\boldsymbol{\theta}) &:= \ell(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}) \\ &= -\frac{1}{2}\Big\{\ln|\boldsymbol{V}(\theta)| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))^{'}\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))\Big\}. \end{aligned} \tag{39}$$

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

With LMMs Restricted (or Residual) maximum likelihood (REML) is a popular alternative to estimate $\boldsymbol{\theta}$. 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 normality [5], and are invariant to the value of $\boldsymbol{\beta}$. The REML estimation of $\boldsymbol{\theta}$ is based on the marginal log–likelihood obtained by integrating out the fixed effects in $\boldsymbol{\beta}$:

$$\ell_r(\boldsymbol{\theta}) \;:=\; \ln\left(\int L(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}\right), \tag{41}$$

where (see Czado (2004))

$$\begin{aligned}\int L(\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta} &= \int \frac{1}{(2\pi)^{N/2}}\,|\boldsymbol{V}(\boldsymbol{\theta})|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right)d\boldsymbol{\beta}\\ &\;\vdots\\ &= \ell_p(\boldsymbol{\theta})-\frac{1}{2}\ln\left|\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\right| + \text{constants}.\end{aligned} \tag{42}$$

2.3.1 Standard errors and inference

Estimation of standard errors. In the marginal model $\boldsymbol{y}\sim N(\boldsymbol{X\beta},\boldsymbol{V}(\boldsymbol{\theta}))$, the covariance of $\hat{\boldsymbol{\beta}}$ in (29) is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) \;=\; (\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X})^{-1}, \tag{43}$$

where $\text{Cov}(\boldsymbol{y}) = \boldsymbol{V}(\boldsymbol{\theta})$ is used. Replacing the unknown $\boldsymbol{\theta}$ with its ML or REML estimate $\hat{\boldsymbol{\theta}}$ and using $\hat{\boldsymbol{V}} := \boldsymbol{V}(\hat{\boldsymbol{\theta}})$, a natural estimate for $\text{Cov}(\hat{\boldsymbol{\beta}})$ is $(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}$. However, this estimate ignores the extra variability originating from the estimation of $\boldsymbol{\theta}$. 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 Chapter XXX where we demonstrate a Bayesian analysis of a Generalized Linear Mixed Model).

The covariance of the empirical BLUP in (32) is equal to

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{u}}) &= \text{Cov}(\boldsymbol{DZ}'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}))\\ &= \boldsymbol{DZ}'\left\{\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right\}\boldsymbol{ZD}.\end{aligned} \tag{44}$$

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

$$\hat{\sigma}^2_{ML} \;=\; \frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2,\quad \hat{\sigma}^2_{REML} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2, \tag{40}$$

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

However, the estimator in (44) ignores the variability in the random vector $\boldsymbol{u}$. Therefore, as suggested by Laird and Ware (1982), inference for $\boldsymbol{u}$ is usually based on $\text{Cov}(\hat{\boldsymbol{u}} - \boldsymbol{u})$. Estimates of the precision of other predictors involving $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ are based on

$$\text{Cov}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} - \boldsymbol{u} \end{bmatrix}, \tag{45}$$

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.

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

$$\begin{aligned} H_0 : \boldsymbol{C\beta} &= \boldsymbol{\zeta} \\ \text{versus } H_1 : \boldsymbol{C\beta} &\neq \boldsymbol{\zeta}. \end{aligned} \tag{46}$$

The Wald test statistic

$$[\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\zeta}]^{'}[\boldsymbol{C}\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^{'}][\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\zeta}] \tag{47}$$

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

$$-2[\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}) - \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})], \tag{48}$$

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

Testing the necessity of random effects requires a hypothesis test involving the variance components. For example, in the varying intercepts model from (7), we want to investigate whether the intercepts of different subjects are significantly different. This corresponds with

$$H_0 : \ \sigma_2^2 = 0 \quad \text{versus} \quad H_1 : \ \sigma_2^2 > 0. \tag{49}$$

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^2_1 distribution, but with a mixture $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. When testing a hypothesis involving s fixed effects parameters and one variance component, the reference distribution is $\frac{1}{2}\chi^2_s + \frac{1}{2}\chi^2_{s+1}$. When more variance components are involved, the complexity of this problem increases, see Ruppert et al. (2003) and related work from these authors.

3 Examples

3.1 Workers' compensation insurance losses

We analyze the data from Illustration 2 on losses observed for workers' compensation insurance risk classes. Variable of interest is Loss_{ij} observed per risk class i and year j. The distribution of the losses is right skewed, which motivates the use of $\log(\mathsf{Loss}_{ij})$ as response variable. To enable out-of-sample predictions, we split the data set in a training (without Loss_{i7}) versus test set (the Loss_{i7} observations). We remove observations corresponding with zero payroll from the data set. Models are estimated on the training set, and centering of covariate Year is applied. Throughout our analysis we include $\log(\mathsf{Payroll})_{ij}$ as an offset in the regression models, since losses should be interpreted relative to the size of the risk class.

Complete pooling. We start with the *'complete pooling'* model, introduced in (1). The model ignores the clustering of data in risk classes and fits an overall intercept (β_0) and an overall slope (β_1) for the effect of Year.

$$\begin{aligned} \log(\mathsf{Loss}_{ij}) &= \log(\mathsf{Payroll}_{ij}) + \beta_0 + \beta_1 \mathsf{Year}_{ij} + \epsilon_{ij} & (50) \\ \epsilon_{ij} &\sim N(0, \sigma^2_\epsilon) \quad i.i.d. & (51) \end{aligned}$$

We fit the model with `lm` in `R`.

```
>fitglm.CP <- lm(log(loss)~yearcentr, offset=log(payroll),data=wclossFit)
>summary(fitglm.CP)

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -4.34023    0.04105 -105.733   <2e-16 ***
yearcentr    0.03559    0.02410    1.477     0.14
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.062 on 667 degrees of freedom
Multiple R-squared:  0.7282,    Adjusted R-squared:  0.7278
F-statistic:  1787 on 1 and 667 DF,  p-value: < 2.2e-16
```

According to this `R` output $\hat{\beta}_0 = -4.34$ (with s.e. 0.041), $\hat{\beta}_1 = 0.036$ (with s.e. 0.024) and $\hat{\sigma}_\epsilon = 1.062$.

No pooling. The fixed effects linear regression model in (2) estimates an intercept for each of the 118 risk classes in the data set. According to model equation (52), the intercepts $\beta_{0,i}$ are unknown, but fixed, whereas the error terms ϵ_{ij} are stochastic.

$$\begin{aligned} log(\mathsf{Loss}_{ij}) &= \log(\mathsf{Payroll}_{ij}) + \beta_{0,i} + \beta_1 \mathsf{Year}_{ij} + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2_\epsilon) \quad i.i.d. & (52) \end{aligned}$$

We fit this model in R by identifying the risk class variable as a factor variable.

```
>fitglm.NP <- lm(log(loss)~0+yearcentr+factor(riskclass), offset=log(payroll),
                 data=wclossFit)
>summary(fitglm.NP)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
yearcentr             0.03843    0.01253   3.067  0.00227 **
factor(riskclass)1   -3.49671    0.22393 -15.615  < 2e-16 ***
factor(riskclass)2   -3.92231    0.22393 -17.516  < 2e-16 ***
factor(riskclass)3   -4.48135    0.22393 -20.012  < 2e-16 ***
factor(riskclass)4   -4.70981    0.22393 -21.032  < 2e-16 ***
...
Residual standard error: 0.5485 on 550 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9983
F-statistic:  3297 on 119 and 550 DF,  p-value: < 2.2e-16
```

The null hypothesis of equal intercepts, $H_0 : \beta_{0,1} = \beta_{0,2} = \ldots = \beta_{0,118} = \beta_0$, is rejected (with p–value < 0.05). Therefore, the *'no pooling'* model significantly improves the *'complete pooling'* model.

```
> anova(fitglm.CP,fitglm.NP)
Analysis of Variance Table

Model 1: log(loss) ~ yearcentr
Model 2: log(loss) ~ 0 + yearcentr + factor(riskclass)
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    667 751.90
2    550 165.48 117    586.42 16.658 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Figure 6 (left) shows the estimates $\hat{\beta}_{0,i}$, plus/minus one standard error, against the size (on log–scale) of the risk class. The size of a risk class is here defined as $\sum_{j=1}^{6} \mathsf{Payroll}_{ij}$. The *'no pooling'* model estimates risk class specific intercepts with reasonable precision.

Linear mixed models: random intercepts. A linear mixed model with random risk class specific intercepts is a meaningful alternative for the *'no pooling'* model in (52). The regression equation is

$$\begin{aligned} \log(\mathsf{Loss}_{ij}) &= \log(\mathsf{Payroll}_{ij}) + \beta_0 + u_{0,i} + \beta_1 \mathsf{Year}_{ij} + \epsilon_{ij} \\ u_{0,i} &\sim N(0, \sigma_u^2) \quad i.i.d. \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \quad i.i.d. \end{aligned} \tag{53}$$

Random intercepts $u_{0,i}$ are independent across risk classes, and independent of the error terms ϵ_{ij}. In R we use the lme4 package to fit this linear mixed model. The package uses REML by default. Results with ML follow by adding REML=FALSE in the lmer(...) statement.

```
> lmm1 <- lmer(log(loss) ~ (1|riskclass)+yearcentr+offset(log(payroll)),
               data=wclossFit)
> print(lmm1)
Linear mixed model fit by REML
Formula: log(loss) ~ (1 | riskclass) + yearcentr + offset(log(payroll))
   Data: wclossFit
  AIC  BIC logLik deviance REMLdev
 1448 1466 -720.2     1431    1440
Random effects:
 Groups    Name        Variance Std.Dev.
 riskclass (Intercept) 0.88589  0.94122
 Residual              0.30145  0.54904
Number of obs: 669, groups: riskclass, 118

Fixed effects:
            Estimate Std. Error t value
(Intercept) -4.31959    0.08938  -48.33
yearcentr    0.03784    0.01253    3.02

Correlation of Fixed Effects:
          (Intr)
yearcentr 0.001
```

The R output shows the following parameter estimates: $\hat{\beta}_0 = -4.32$ (s.e. 0.089), $\hat{\beta}_1 = 0.037$ (s.e. 0.013), $\hat{\sigma}_u = 0.94$ and $\hat{\sigma}_\epsilon = 0.55$. In Figure 6 (right) we plot the point predictions for the $u_{i,0}$'s, and their corresponding standard errors, against size of the risk class. To create this plot we refit the linear mixed model and do not include an intercept.

The point estimates of the random intercepts obtained with the *'no pooling'* model in (52) and the linear mixed model in (53) are similar in this example. For the standard errors of the random intercepts in the LMM we use the following instructions

```
str(rr1 <- ranef(lmm0, condVar = TRUE))
my.se.risk = sqrt(as.numeric(attributes(rr1$riskclass)$postVar)),
```

which calculates the variance of $\boldsymbol{u}|\boldsymbol{y}$ (see XXX and the footnote below (30)), conditional on the maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Thus, these standard errors are different from the approach outlined in (44). We are aware of the fact that they do not account for all sources of variability involved.

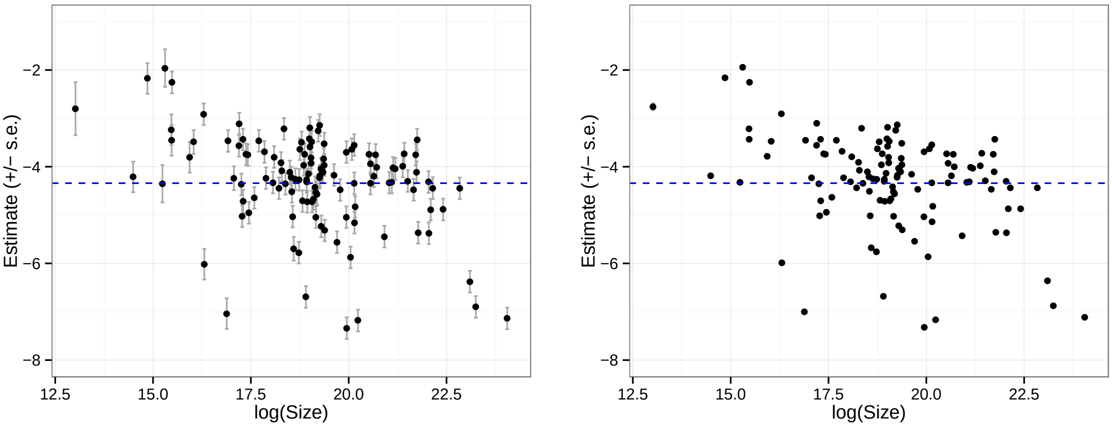

Figure 6: *Point estimates for risk class specific intercepts, plus/minus one standard error. Results from no pooling approach (left) and linear mixed model (right). The dashed line is $y = -4.34$, i.e. the overall intercept from the complete pooling model.*

Linear mixed models: random intercepts and slopes. We now extend the LMM in (53) and allow for random slopes as well as random intercepts. This is an example of the 'multiple random effects per level' setting from Section 2.2. The model equation is

$$\begin{aligned} \log(\mathsf{Loss}_{ij}) &= \log(\mathsf{Payroll}_{ij}) + \beta_0 + u_{0,i} + \beta_1 \mathsf{Year}_{ij} + u_{1,i}\mathsf{Year}_{ij} + \epsilon_{ij}, \\ \boldsymbol{u}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{\theta})) \quad i.i.d. \\ \epsilon_{ij} &\sim N(0, \sigma^2_{\epsilon}) \quad i.i.d. \end{aligned} \tag{54}$$

The random effects vector $\boldsymbol{u}_i$ is now bivariate, say with $\text{Var}(u_{i,0}) = \theta_0$, $\text{Var}(u_{i,1}) = \theta_1$ and $\text{Cov}(u_{i,0}, u_{i,1}) = \theta_{01}$. Random effects are independent across risk classes, and independent of the error terms ϵ_{ij}. We fit this model with `lmer` as follows.

```
> lmm2 <- lmer(log(loss) ~ (1+yearcentr|riskclass)+yearcentr+offset(log(payroll)),
          data=wclossFit)
> print(lmm2)
Linear mixed model fit by REML
Formula: log(loss) ~ (1 + yearcentr | riskclass) + yearcentr + offset(log(payroll))
   Data: wclossFit
  AIC  BIC logLik deviance REMLdev
 1451 1478 -719.4     1429    1439
Random effects:
 Groups    Name        Variance Std.Dev. Corr
 riskclass (Intercept) 0.885937 0.941242
           yearcentr   0.003171 0.056312 -0.195
 Residual              0.290719 0.539184
Number of obs: 669, groups: riskclass, 118
```

```
Fixed effects:
            Estimate Std. Error t value
(Intercept) -4.32030    0.08929  -48.38
yearcentr    0.03715    0.01340    2.77

Correlation of Fixed Effects:
          (Intr)
yearcentr -0.072
```

In this output $\hat{\theta}_0 = 0.89$, $\hat{\theta}_1 = 0.0032$ and $\hat{\theta}_{01} = -0.010$. We test whether the structure of random effects should be reduced, i.e. $H_0 : \theta_1 = 0$ (with θ_1 the variance of random slopes), using an `anova` test comparing models (53) and (54).

```
> anova(lmm1,lmm2)
Data: wclossFit
Models:
lmm1: log(loss) ~ (1 | riskclass) + yearcentr + offset(log(payroll))
lmm2: log(loss) ~ (1 + yearcentr | riskclass) + yearcentr + offset(log(payroll))
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
lmm1  4 1438.5 1456.6 -715.27
lmm2  6 1440.9 1468.0 -714.46 1.6313      2     0.4423
```

When performing the corresponding LRT the software automatically refits `lmm1` and `lmm2` with ML (instead of REML), as required (see our discussion in Section 2.3.1). This explains why the `AIC`, `BIC` and `logLik` values differ from those printed above. The observed `Chisq` test statistic and reported p–value indicate that $H_0 : \sigma_1^2 = 0$ can not be rejected. The model with only random intercepts is our preferred specification.

Out–of–sample predictions. We compare out–of–sample predictions of Loss_{i7}, for given $\mathsf{Payroll}_{i7}$, as obtained with models (50), (52) and (53). Figure 7 plots observed versus fitted losses (on log scale) for (from left to right) the *complete pooling*, the random intercepts and the *no pooling* linear regression model.

3.2 Hachemeister data

We present an analysis of the Hachemeister data using three simple linear mixed models. Chapter XXX presents an in depth discussion of credibility models for this data set (namely the Bühlmann, Bühlmann–Straub and Hachemeister credibility models). By combining the `R` scripts prepared for our illustration with the scripts from Chapter XXX, readers obtain relevant illustrations of credibility models in `R` and their analogue interpretation as LMMs.

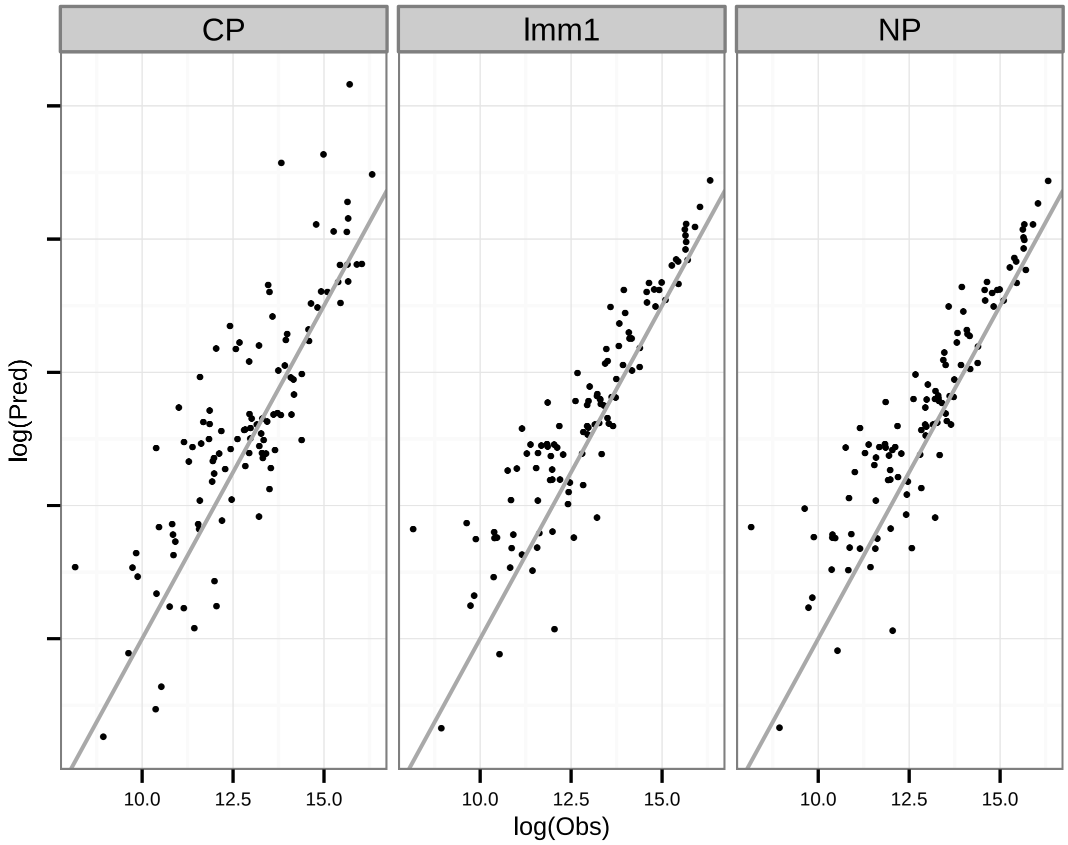

Figure 7: *Out–of–sample predictions for* Loss_{i7} *versus observed losses, as obtained with model* (50) *('CP', complete pooling),* (53) *('lmm1', random intercepts) and* (52) *('NP', no pooling): losses on workers' insurance compensation.*

Random intercepts, no weights. Response variable is the average loss per claim (i.e. Ratio_{ij}), per state i ($i = 1, \ldots, 5$) and quarter j ($j = 1, \ldots, 12$). A basic random state intercept model for Ratio_{ij} is

$$\begin{aligned} \mathsf{Ratio}_{ij} &= \beta_0 + u_{i,0} + \epsilon_{ij} \\ u_{i,0} &\sim N(0, \sigma_u^2) \quad i.i.d. \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \quad i.i.d. \end{aligned} \tag{55}$$

Apart from the normality assumption, actuaries recognize the so–called Bühlmann credibility model, as Chapter XXX explains.

Random intercepts, including weights. Our response variable is average loss per claim, constructed as total loss (per state and quarter) divided by the corresponding number of claims. This average loss is more precise when more claims have been observed.

We therefore include the number of observed claims as weights (w_{ij}) in our LMM.

$$\begin{aligned} \mathsf{Ratio}_{ij} &= \beta_0 + u_{i,0} + \epsilon_{ij} \\ u_{i,0} &\sim N(0, \sigma_u^2) \quad i.i.d. \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2 / w_{ij}) \quad i.i.d. \end{aligned} \tag{56}$$

The model equation and variance assumptions (apart from normality) correspond with the Bühlmann–Straub credibility model. Including weights goes as follows in `R` `lme4`:

```
> lmmBS <- lmer(ratio ~ (1|state),weights=weight,data=hach)
> print(lmmBS)
Linear mixed model fit by REML
Formula: ratio ~ (1 | state)
   Data: hach
  AIC  BIC logLik deviance REMLdev
 1301 1307 -647.5     1306    1295
Random effects:
 Groups   Name        Variance  Std.Dev.
 state    (Intercept)    22.326   4.725
 Residual             47928.954 218.927
Number of obs: 60, groups: state, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept) 1688.934      2.265   745.6
```

The risk (or: credibility) premium for state i is $\hat{\beta}_0 + \hat{u}_{i,0}$, and is available in `R` as follows

```
## get fixed effects
fe <- fixef(lmmBS)
## get random intercepts
re <- ranef(lmmBS)
## calculate credibility premiums in this lmm
pred.lmmBS <- fe[1]+re$state
> t(pred.lmmBS)
                  1        2        3        4        5
(Intercept) 2053.18 1528.509 1790.053 1468.113 1604.815
```

Chapter XXX illustrates how traditional actuarial credibility calculations are available in the `actuar` package in `R`. The credibility premiums obtained with Bühlmann–Straub are close to – but not exactly the same as – the premiums obtained with (56). Note that the actuarial credibility calculations use method of moments for parameter estimation, whereas our LMMs use (RE)ML.

```
> ## BS model (Buhlmann-Straub credibility model)
> ## use actuar package, and hachemeister data as available in this package
> fitBS <- cm(~state, hachemeister,ratios = ratio.1:ratio.12,
                 weights = weight.1:weight.12)
> pred.BS <- predict(fitBS) # credibility premiums
> pred.BS
[1] 2055.165 1523.706 1793.444 1442.967 1603.285
```

Random intercepts and slopes, including weights. We extend the random intercepts model to a random intercepts and slopes model, using the period of observation as regressor.

$$\begin{aligned}
\mathsf{Ratio}_{ij} &= \beta_0 + u_{i,0} + \beta_1 \mathsf{period}_{ij} + u_{i,1}\mathsf{period}_{ij} + \epsilon_{ij} \\
\boldsymbol{u}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{\theta})) \quad \textit{i.i.d.} \\
\epsilon_{ij} &\sim N(0, \sigma^2_\epsilon / w_{ij}) \quad \textit{i.i.d.}
\end{aligned} \tag{57}$$

Our analysis uses period_{ij} as the quarter ($j = 1, \ldots, 12$) of observation. The use of a centered version of period is discussed in Chapter XXX. In R the (1+period|state) instruction specifies random intercepts and slopes per state.

```
> lmmHach <- lmer(ratio ~ period+(1+period|state),weights=weight,data=hach)
> lmmHach
Linear mixed model fit by REML
Formula: ratio ~ period + (1 + period | state)
   Data: hach
  AIC  BIC logLik deviance REMLdev
 1242 1255 -615.1     1247    1230
Random effects:
 Groups   Name        Variance   Std.Dev.  Corr
 state    (Intercept) 4.1153e+00   2.02863
          period      1.9092e-01   0.43695 1.000
 Residual             1.6401e+04 128.06735
Number of obs: 60, groups: state, 5

Fixed effects:
             Estimate Std. Error t value
(Intercept) 1501.5452     1.1265  1333.0
period        27.7333     0.2172   127.7

Correlation of Fixed Effects:
       (Intr)
period 0.540
```

Using LMM (57) the state specific risk premium for the next time period, is

$$\widehat{E[\mathsf{Ratio}_{i,13}|\boldsymbol{u}_i]} = \hat{\beta}_0 + \hat{u}_{i,0} + \hat{\beta}_1 \cdot 13 + \hat{u}_{i,1} \cdot 13 \tag{58}$$

```
> t(pred.lmmHach)
         [,1]     [,2]     [,3]     [,4]    [,5]
[1,] 2464.032 1605.676 2067.279 1453.923 1719.48.
```

These premiums correspond with the results (obtained with SAS) reported in Frees et al. (1999) (see Table 3, columns 'Prediction and standard errors'). These authors also investigate linear mixed models as a user friendly and computationally attractive alternative for actuarial credibility models. The traditional Hachemeister credibility premiums are available in R as follows (see also the 'Base' results in Table 3 from Frees et al. (1999))

```
fitHach <- cm(~state, hachemeister,regformula = ~time, regdata =
          data.frame(time = 1:12),ratios = ratio.1:ratio.12,
          weights = weight.1:weight.12)
pred.Hach <- predict(fitHach, newdata = data.frame(time = 13))
# cred.premium
# > pred.Hach
# [1] 2436.752 1650.533 2073.296 1507.070 1759.403
```

Once again, with linear mixed models we obtain premiums that are close to, but do not replicate, the traditional actuarial credibility results. Differences in parameter estimation techniques explain why these results are not identical.

Using a LRT we verify whether model (57) should be reduced to the model with random intercepts only. The *p*–value indicates that this is not case.

```
lmmHach2 <- lmer(ratio ~ period+(1|state),weights=weight,data=hach)
anova(lmmHach,lmmHach2)
#Data: hach
#Models:
#lmmHach2: ratio ~ period + (1 | state)
#lmmHach: ratio ~ period + (1 + period | state)
#         Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
#lmmHach2  4 1272.2 1280.5 -632.08
#lmmHach   6 1258.7 1271.2 -623.32 17.521      2  0.0001568 ***
#---
#Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Figure 8 illustrates the fit of a *complete pooling* (dark grey, dashed line), a *no pooling* (black, dashed line) and a LMM with random intercepts and slopes (black, solid line). The regression equations for the *complete* and *no* pooling model are

$$\begin{aligned} \mathsf{Ratio}_{ij} &= \beta_0 + \beta_1 \mathsf{period}_{ij} + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2_{\epsilon}/w_{ij}), \end{aligned} \tag{59}$$

and

$$\begin{aligned} \mathsf{Ratio}_{ij} &= \beta_{0,i} + \beta_{1,i}\mathsf{period}_{ij} + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2/w_{ij}), \end{aligned} \tag{60}$$

respectively.

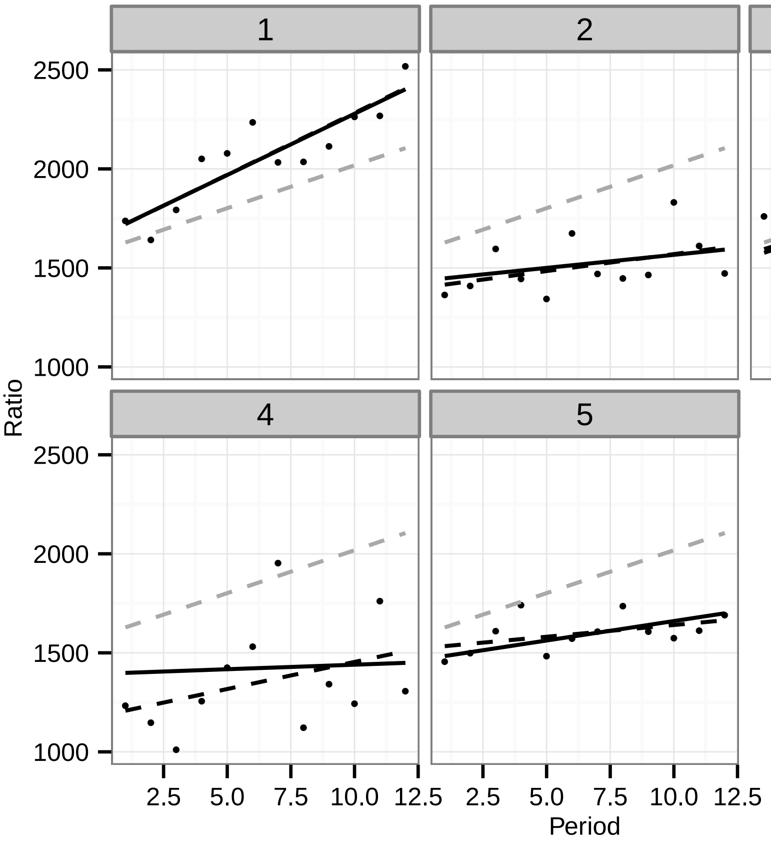

Figure 8: *Fit of a complete pooling (dark grey, dashed line), a no pooling (black, dashed line) and a LMM with random intercepts and slopes (black, solid line): Hachemeister data (no centering of period).*

3.3 Credit insurance data

We analyze the data from Illustration 5 and demonstrate the use of crossed random effects (see Section 2.2) with `lme4`. The response variable of interest is $\mathsf{Payment}_{ijt}$, where $i = 1, 2, 3$ denotes status and $j = 1, 2, 3$ is for working experience of the insured, t is an index going over all observation in cell (i, j). Dannenburg et al. (1996) use these data to demonstrate the principles of a so–called cross classification credibility model, with model equation (in typical actuarial credibility notation)

$$\mathsf{Payment}_{ijt} = m + \Xi_i^{(1)} + \Xi_j^{(2)} + \Xi_{ij}^{(12)} + \Xi_{ijt}^{(123)}. \tag{61}$$

Hereby, m is an overall intercept, $\Xi_i^{(1)}$ is a random effect for level i in factor (1) (i.e. status), $\Xi_j^{(2)}$ a random intercept for level j in factor (2) (i.e. experience) and $\Xi_{ij}^{(12)}$ is a random effect for the interaction of level i and j. $\Xi_{ijt}^{(123)}$ is an error term for observation t from the combined level i and j. Dannenburg et al. (1996) obtain the following credibility premiums

	experience		
status	1	2	3
1	181.05	238.18	277.77
2	172.11	229.16	268.8
3	225.29	282.24	323.68

Table 2: *Credibility premiums obtained with crossed classification credibility model per combination of* status *and* experience *risk class: credit insurance data.*

The analysis of this data by means of a linear mixed model with crossed random effects (i.e. `(1 | status:experience)`), is directly available in `R`.

$$\begin{aligned} \mathsf{Payment}_{ijt} &= m + u_i^{(1)} + u_j^{(2)} + u_{ij}^{(12)} + \epsilon_{ijt} \\ u_i^{(1)} &\sim N(0, \sigma_1^2) \\ u_j^{(2)} &\sim N(0, \sigma_2^2) \\ u_{ij}^{(12)} &\sim N(0, \sigma_{12}^2) \\ \epsilon_{ijt} &\sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{62}$$

where i and j run over all levels in factors 1 (status) and 2 (experience) and we assume all random variables to be independent.

```
> lmm2 <- lmer(payment ~ 1+(1|experience)+(1|status)+(1|status:experience)
               ,data=credit)
> print(lmm2)
Linear mixed model fit by REML
Formula: payment ~ 1 + (1 | experience) + (1 | status) + (1 | status:experience)
   Data: credit
  AIC  BIC logLik deviance REMLdev
 5241 5261  -2616     5240    5231
Random effects:
 Groups            Name        Variance  Std.Dev.
 status:experience (Intercept)    14.611   3.8224
 status            (Intercept)   992.791  31.5086
 experience        (Intercept)  2569.330  50.6886
 Residual                      26990.398 164.2875
Number of obs: 401, groups: status:experience, 9; status, 3; experience, 3
```

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)   244.25      35.44   6.892
```

The resulting risk premiums as obtained with `lme4` are very close to the credibility premiums in Table 2.

```
      experience
status        1        2        3
     1 181.0253 238.1813 277.7692
     2 172.1086 229.1551 268.7954
     3 225.2921 282.2424 323.6784
```

Our analysis directly uses Payment as response variable to facilitate the comparison between the credibility and linear mixed model calculations. However, the positivity and right skewness of Payment suggests the use of a lognormal or gamma distribution for this response.

4 Further readings and illustrations

We recommend Czado (2004), Gelman and Hill (2007), Frees (2004a), McCulloch and Searle (2001), Ruppert et al. (2003) and Verbeke and Molenberghs (2000) as further readings on linear mixed models. The use of LMMs for smoothing purposes is not discussed above, but interested readers can find below a brief introduction and useful references.

Illustration 9 (Smoothing with mixed models)**.** *A semiparametric regression model incorporates both parametric as well as nonparametric functional relationships between a response and a set of covariates. 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 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 linear models.*

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 towards 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*** *(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 problems in predictive modeling.*

For example, Figure 9 shows an application of the P–splines in estimating insurance loss reserves. In this example, the incremental paid insurance losses, represented by the dots in the plot, exhibit a nonlinear dependence upon the report lag (the x-axis). Standard loss reserving methods will specify a model with these lags as categorical covariates. In contrast, P–splines allow us to estimate a smooth functional relationship between paid losses and report lags. One advantage over the reserving model with dummy variables is the reduced number of model parameters because generally a small number of knots can capture the observed pattern sufficiently well. The example shown here is based on a four-knot penalized spline, and Zhang and Dukic (2012) find that the resulting model has significantly better predictive performance than a dummy-variable-based reserving model. Another benefit is that estimates at any time point can be produced based on interpolation or extrapolation of the estimated functional form. This can be very helpful when the goal of a reserving study is to make forecasts for a short period ahead, say one month or a quarter.

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 GAMS / spatial statistics.]

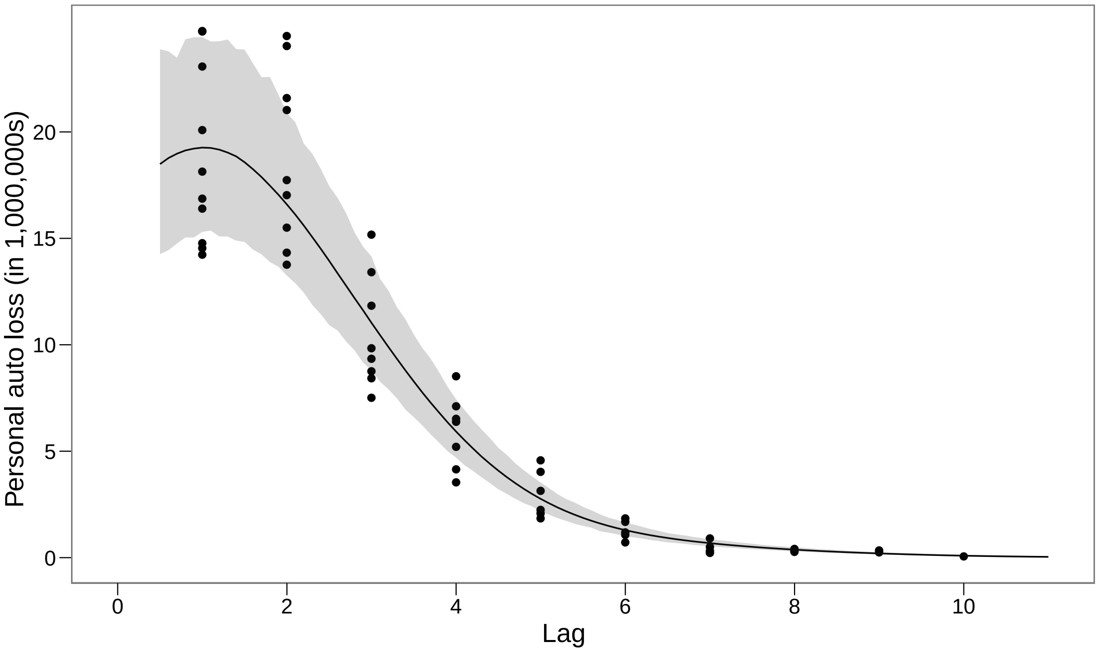

Figure 9: *The plot of the company-level smoother (incremental losses) along with the 50% prediction interval for a loss triangle.*

References

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

Antonio, K. and Beirlant, J. (2008). Issues in claims reserving and credibility: a semiparametric approach with mixed models. *Journal of Risk and Insurance*, 75(3):643–676.

Antonio, K., Frees, E., and Valdez, E. (2010). A multilevel analysis of intercompany claim counts. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 40(1):151–177.

Antonio, K. and Valdez, E. (2012). Statistical aspects of *a priori* and *a posteriori* risk classification in insurance. *Advances in Statistical Analysis*, 96(2):187–224.

Bühlmann, H. and Gisler, A. (2005). *A course in credibility theory and its applications.* Springer Verlag, Berlin.

Czado, C. (2004). *Linear Mixed Models.* Lecture slides on GLM, TU Munchen.

Dannenburg, D., Kaas, R., and Goovaerts, M. (1996). *Practical actuarial credibility models.* Institute of actuarial science and econometrics, University of Amsterdam.

Frees, E. (2004a). *Longitudinal and panel data. Analysis and applications in the social sciences.* Cambridge University Press.

Frees, E. (2004b). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences.* Cambridge University Press, Cambridge.

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

Frees, E., Young, V., and Luo, Y. (2001). Case studies using panel data models. *North American Actuarial Journal*, 5(4):24–42.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435.

Gelman, A. and Hill, J. (2007). *Applied Regression and Multilevel (Hierarchical) Models.* Cambridge University Press, Cambridge.

Hachemeister, C. (1975). *Credibility: Theory and Applications*, chapter 'Credibility for regression models with application to trend', pages 129–163. Academic Press, New York.

Kacker, R. and Harville, D. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79:853–862.

Klugman, S. (1992). *Bayesian statistics in actuarial science with emphasis on credibility.* Kluwer, Boston.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Makov, U., Smith, A., and Liu, Y. (1996). Bayesian methods in actuarial science. *The Statistician*, 45(4):503–515.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear and Mixed Models.* Wiley Series in Probability and Statistics, Wiley, New York.

Robinson, G. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science*, 6:15–51.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression.* Cambridge University Press, Cambridge.

Scollnik, D. (1996). An introduction to Markov Chain Monte Carlo methods and their actuarial applications. *Proceedings of the Casualty Actuarial Society Forum*, LXXXIII:114–165.

Searle, S., Casella, G., and McCulloch, C. (2008). *Variance components.* Wiley.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* Springer Series In Statistics, New York.

Wood, S. (2006). *Generalized Additive Models: An introduction with R.* Chapman & Hall, CRC Texts in Statistical Science.

Zhang, Y. and Dukic, V. (2012). Predicting multivariate insurance loss payments under the bayesian copula framework. *The Journal of Risk and Insurance.* in press, DOI: 10.1111/j.1539-6975.2012.01480.x.

Zhang, Y., Dukic, V., and Guszcza, J. (2012). A bayesian nonlinear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society, Series A*, 175:637–656.

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