Mixed Models

(An introduction)

R. Bellio & N. Torelli

Spring 2018

University of Udine & University of Trieste

Basic ideas

Linear Mixed Models (LMMs): an introductory example

Some theory

Extensions

Basic ideas

Intro: random effects

By **mixed models** we denote statistical models which include regression parameters (such as the usual coefficients predictors) plus **random effects**.

As written in the CS book

> *Random variables in a model that are not associated with the independent random variability of single observations, are termed random effects.*

Random effects provides a more precise description of the stochastic structure in the data. In particular, they are useful for *multiple levels of randomness*, like those arising in *hierarchical data*.

Hierarchical data

Typical examples of hierarchical data include

- Students nested in schools, where we have two kind of statistical units: students and schools, which are both sampled by a well defined population of interest. Here a **multilevel model** is called for.
- Patients treated in different hospitals, like in **multicenter clinical studies**, with variability existing among hospitals and also among patients.
- Repeated measurements made on the several subjects, like in **longitudinal studies** or **panel data**. Here a sensible model must account for the variability related to the errors occurring in each measurement, plus the intrinsic variability due to the different subject.

Mixed models are ubiquitous in statistical applications, and are one of the most important tools for the applied statistician.

Again, we will introduce their precise nature by means of a running example.

Linear Mixed Models (LMMs): an introductory example

An actuarial example: Workers Compensation

Dataframe `WorkersComp` in the R package `insuranceData`:

> *Standard example in worker's compensation insurance, examining losses due to permanent, partial disability claims. We consider n=121 occupation, or risk, classes, over T=7 years.*

Main variables (after some transformations, for a sample size of 778 observations):

- `CLASS` : Occupation class identifier (118 different classes, after data pre-cleaning)
- `YR` : Year identifier, 1-7
- `PR` : Payroll, a measure of exposure to loss, in tens of millions of dollars
- `LOSS` : Losses related to permanent partial disability, in tens of millions of dollars
- `PP` : Pure premium (loss per dollar of payroll)

A longitudinal data set

Most classes have one observation per year; for a few classes less than 7 years are observed.

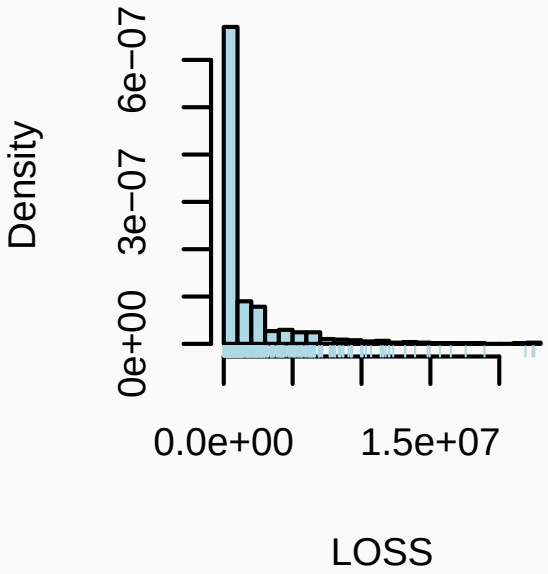
In particular the distribution of classes according to number of years observed is

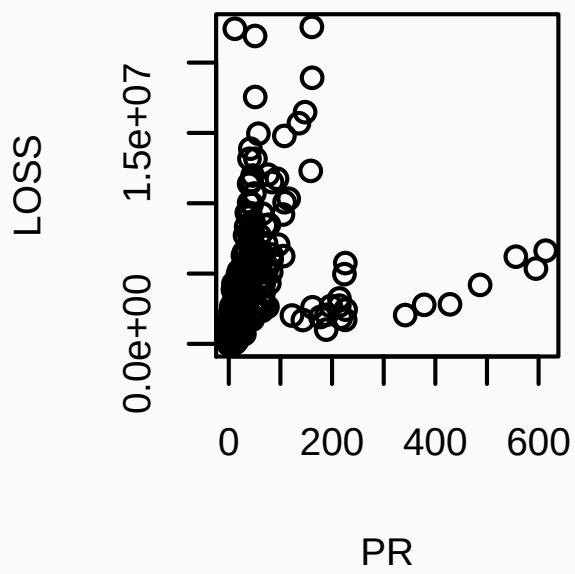
```
##
##  1 year 2 years 3 years 4 years 5 years 6 years 7 years
##       1       1       4       4       1       7     100
```

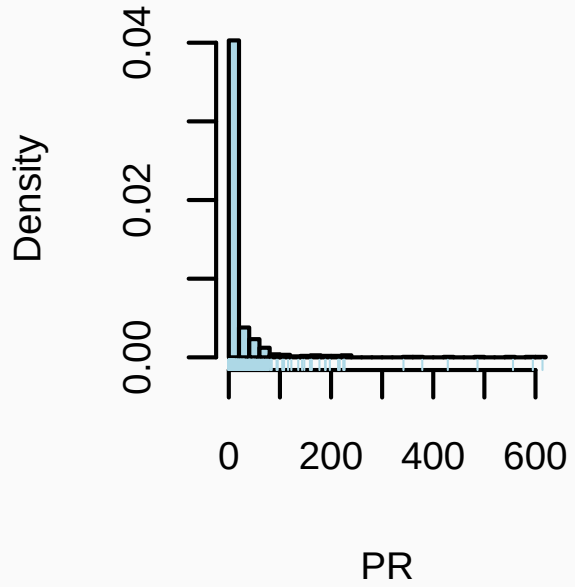
As usual, the first thing to do in any data analysis is to look at the data.

In this case, we shall also try to do this by accounting for **the hierarchical nature of the data set**.

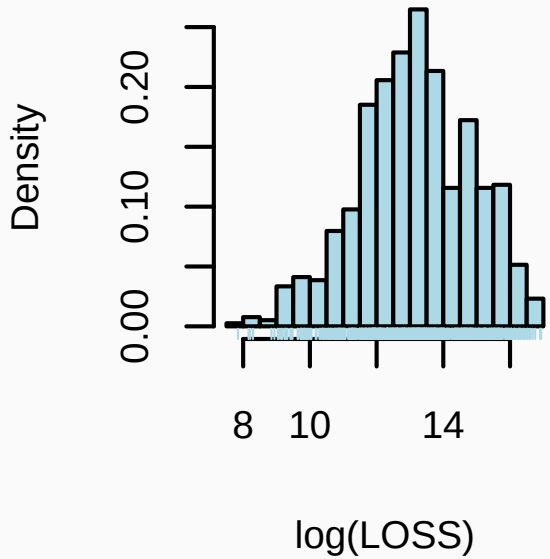
Workers data: distribution of PR and LOSS

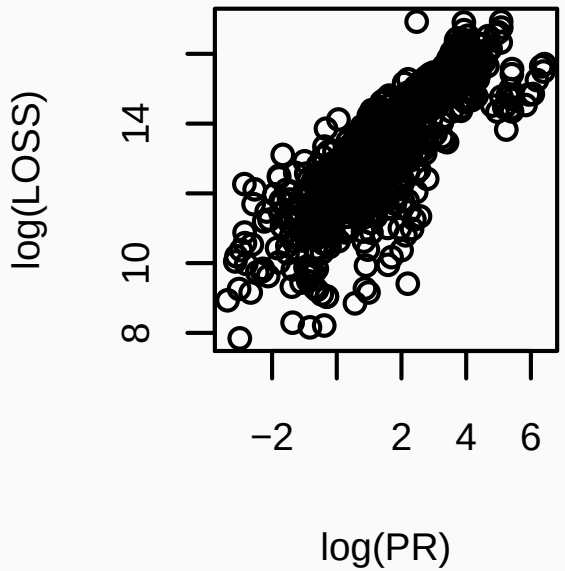


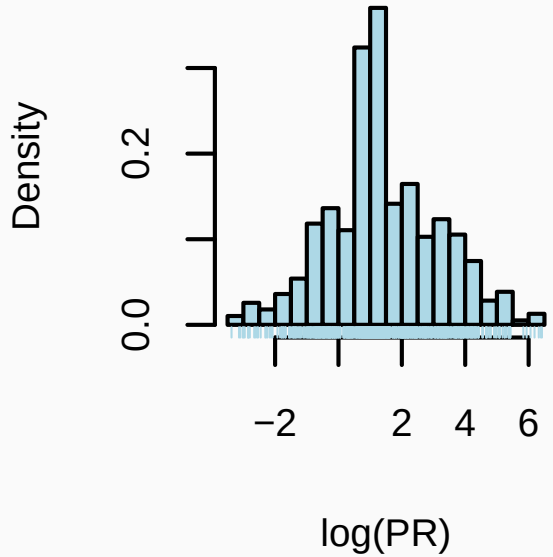




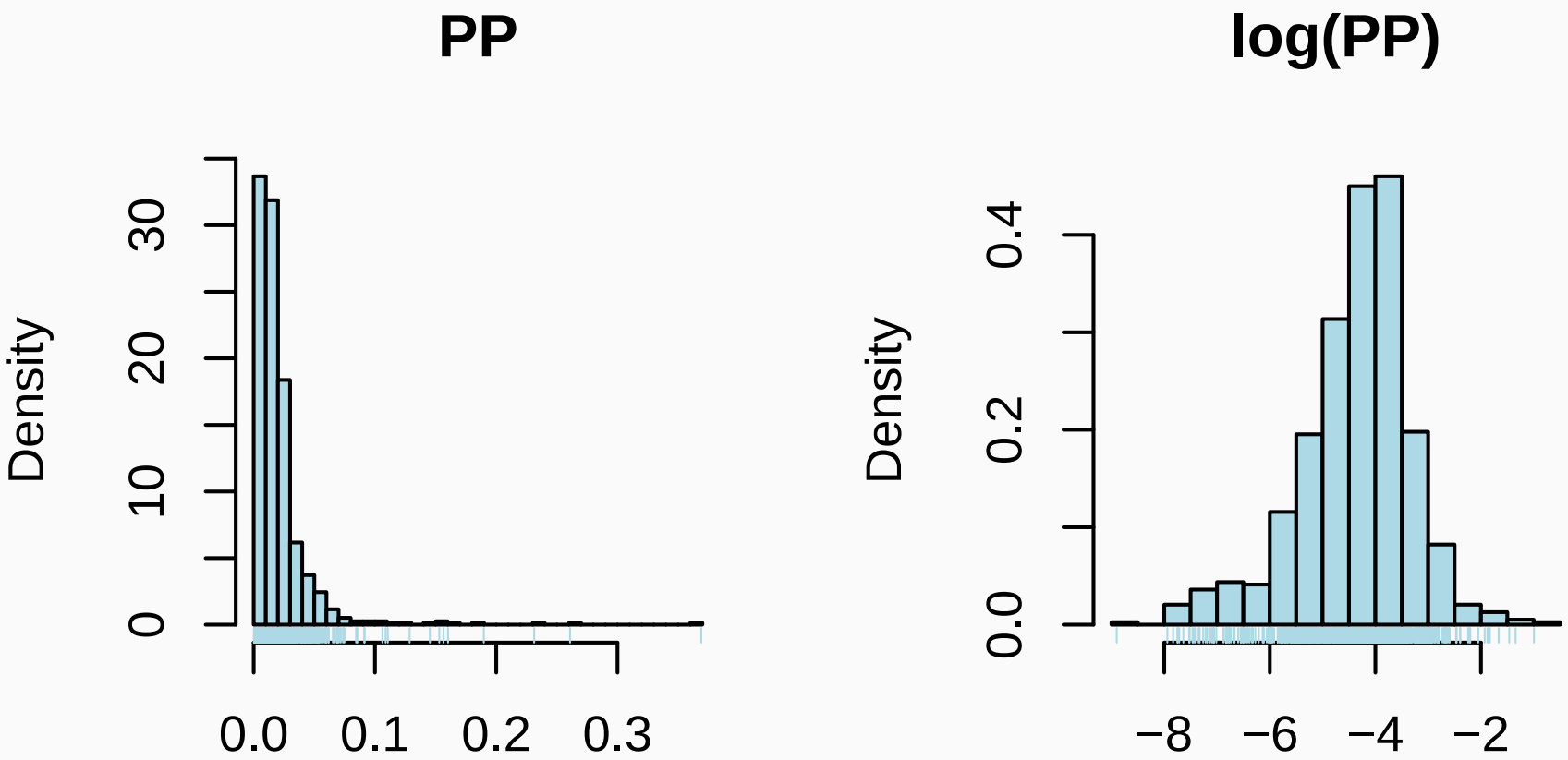
Workers data: distribution of PR and LOSS



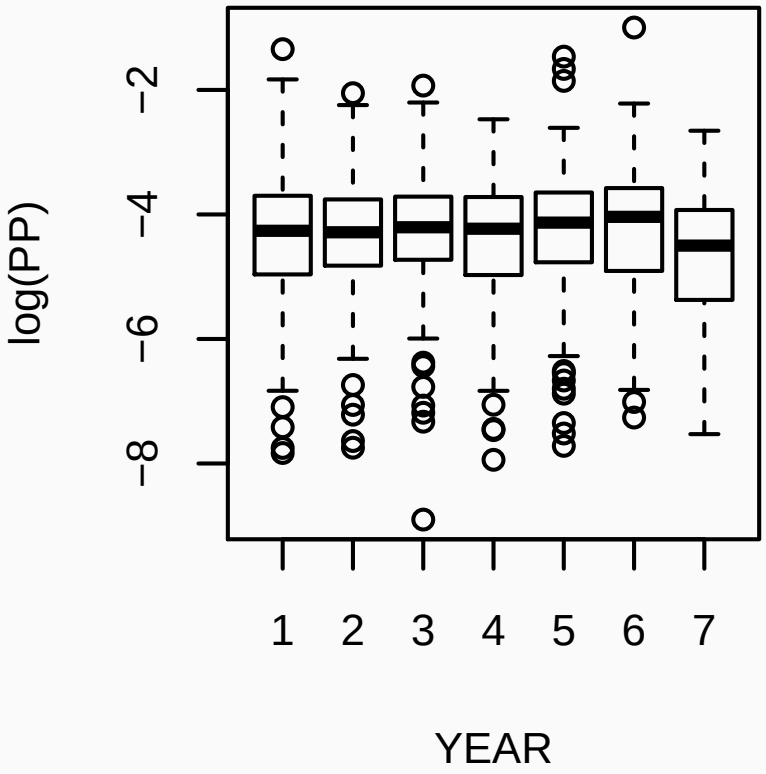


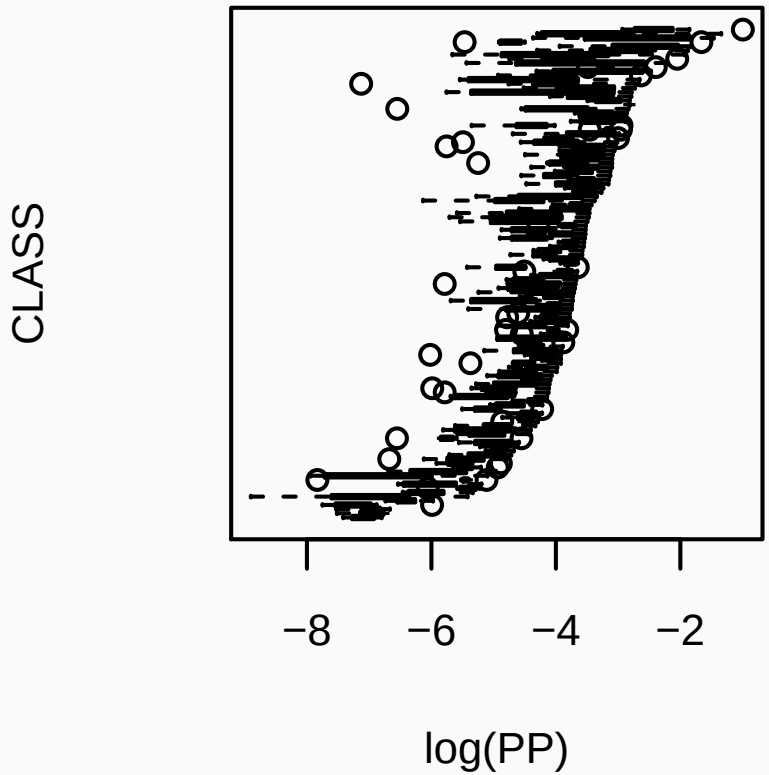


Distribution of PP: univariate analysis

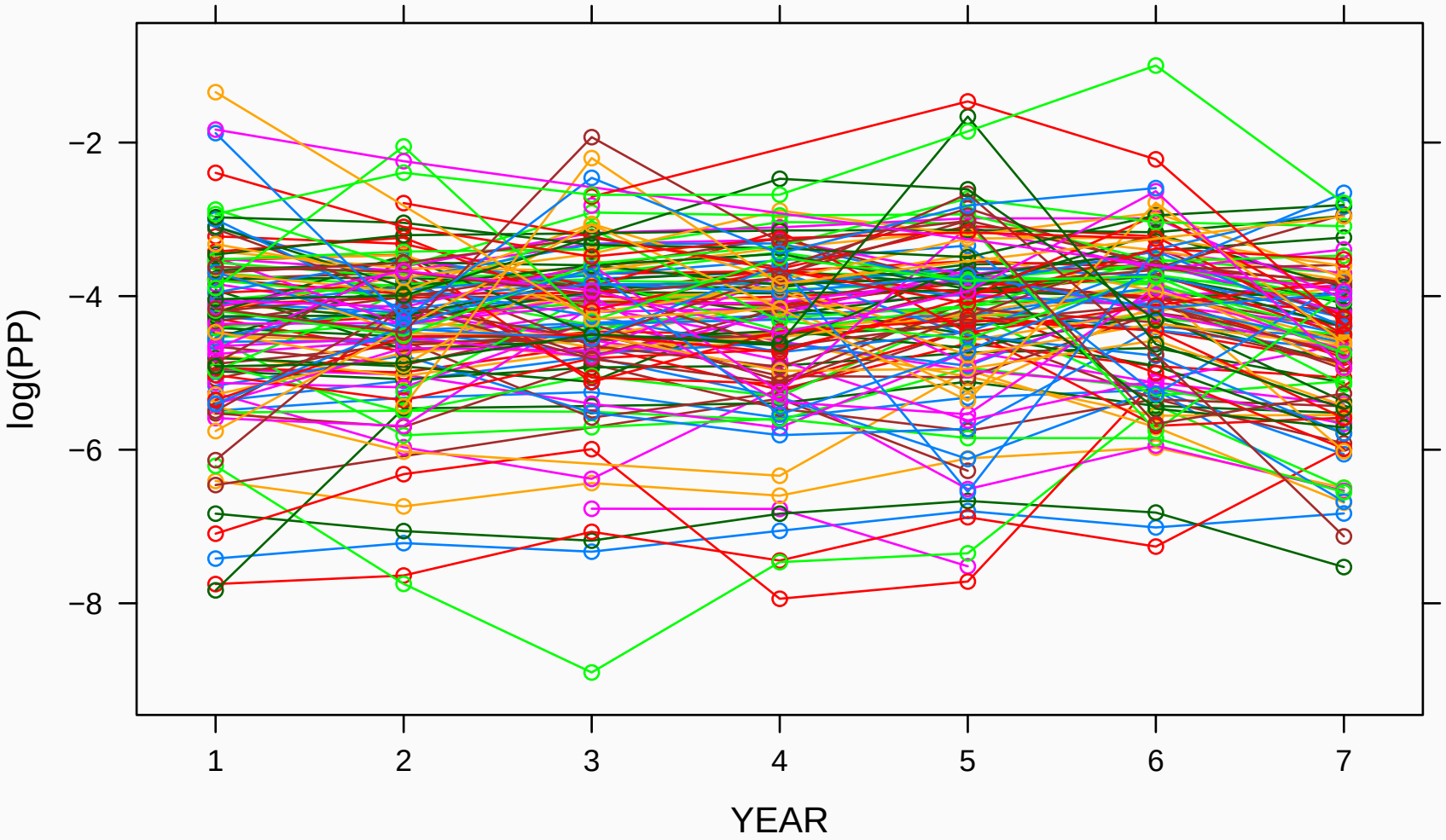


Workers data: log(PP) vs YEAR and CLASS

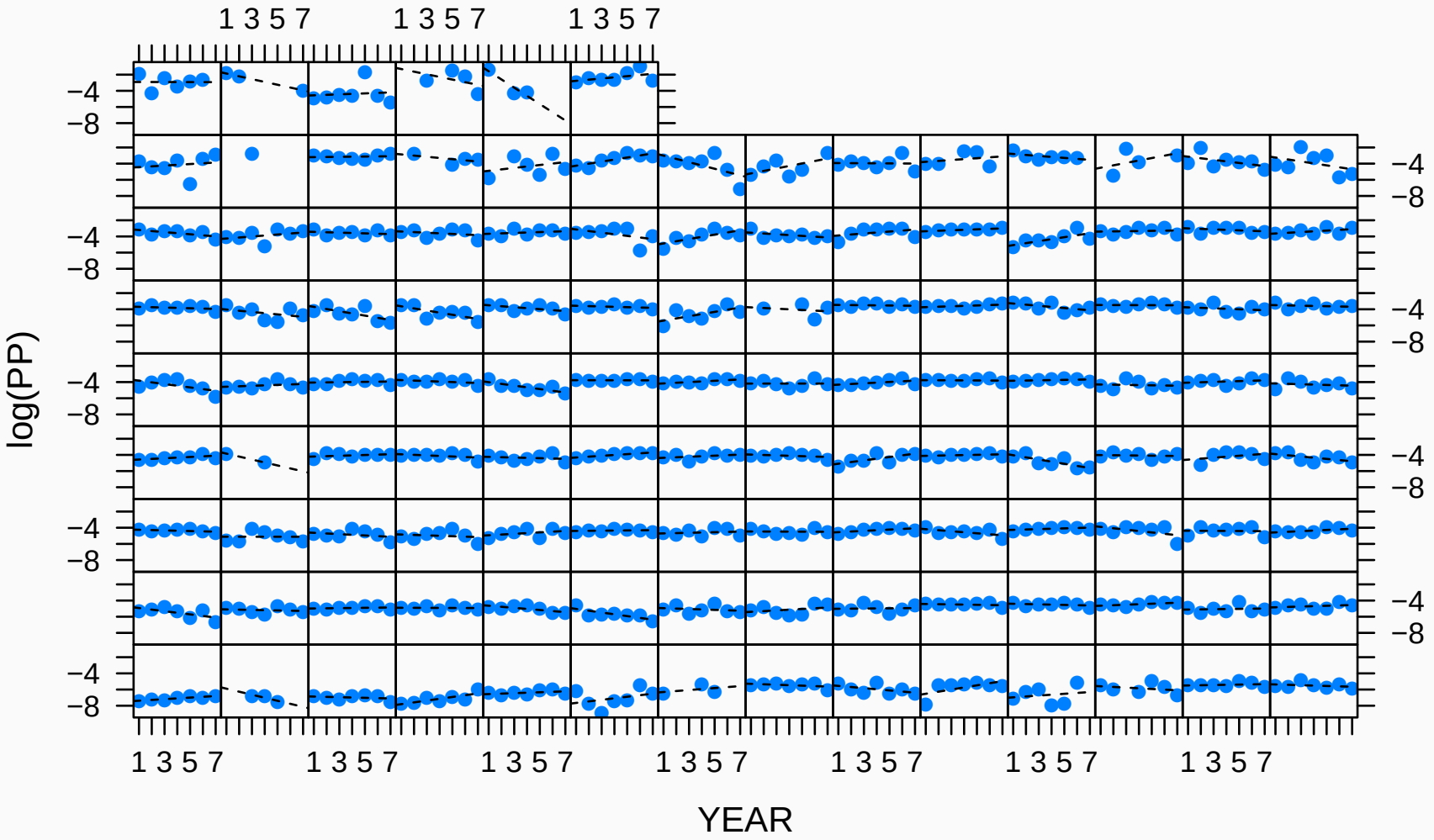




Workers data: log(PP) vs YEAR grouped by CLASS



Workers data: log(PP) vs YEAR grouped by CLASS



Possible simple models for `log(PP)`

The exploratory analysis suggests some alternatives

- Model 1: Simple linear regression of `log(PP)` vs `YEAR`

$$\texttt{log(PP)}_{it} = \beta_0 + \beta_1 \, \texttt{YEAR}_{it} + \varepsilon_{it}$$

  with $i = 1, \ldots, 118$ and $t = 1, \ldots, 7$

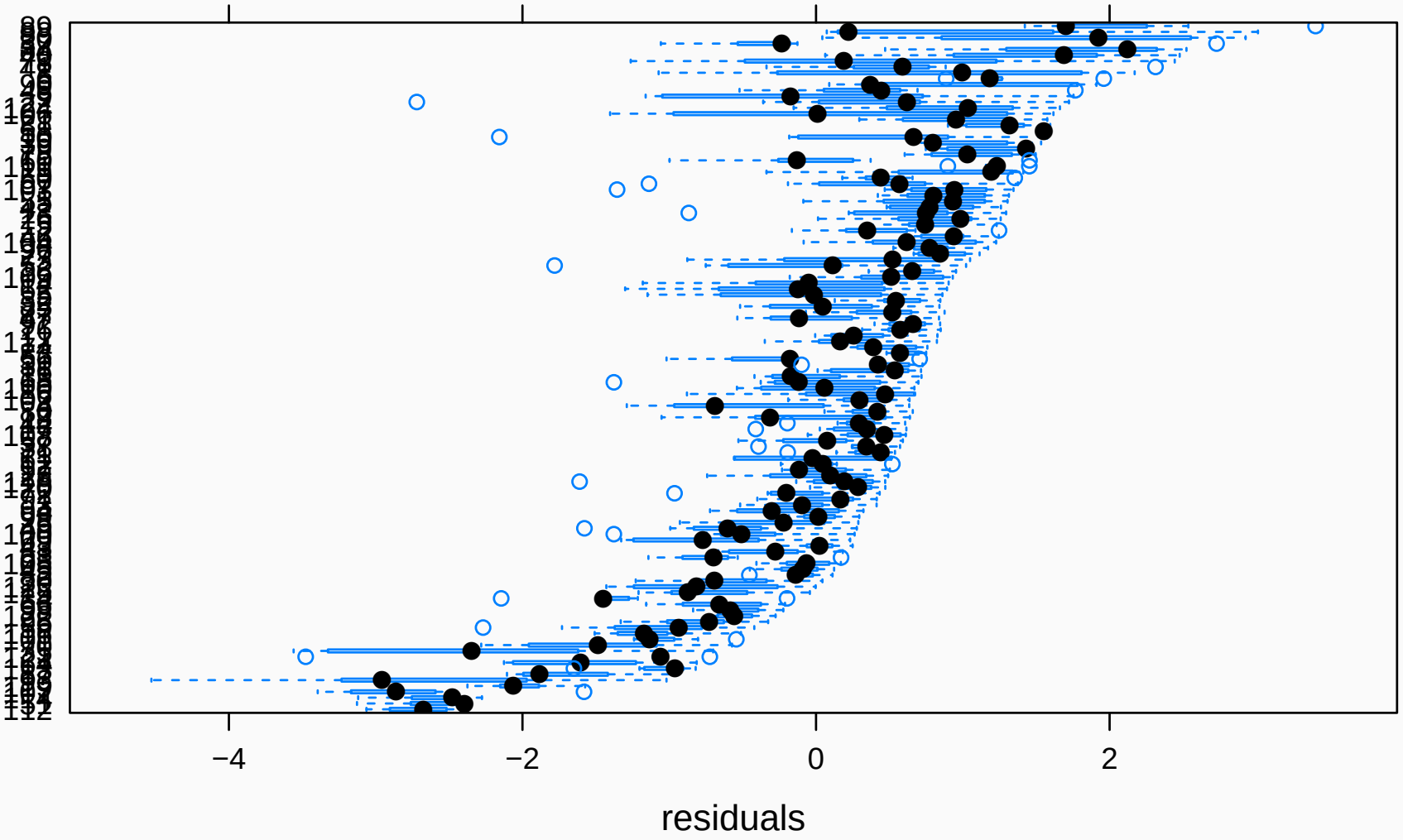
- Model 2: Regression lines with different intercept for each class

$$\texttt{log(PP)}_{it} = \beta_{0i} + \beta_1 \, \texttt{YEAR}_{it} + \varepsilon_{it}$$

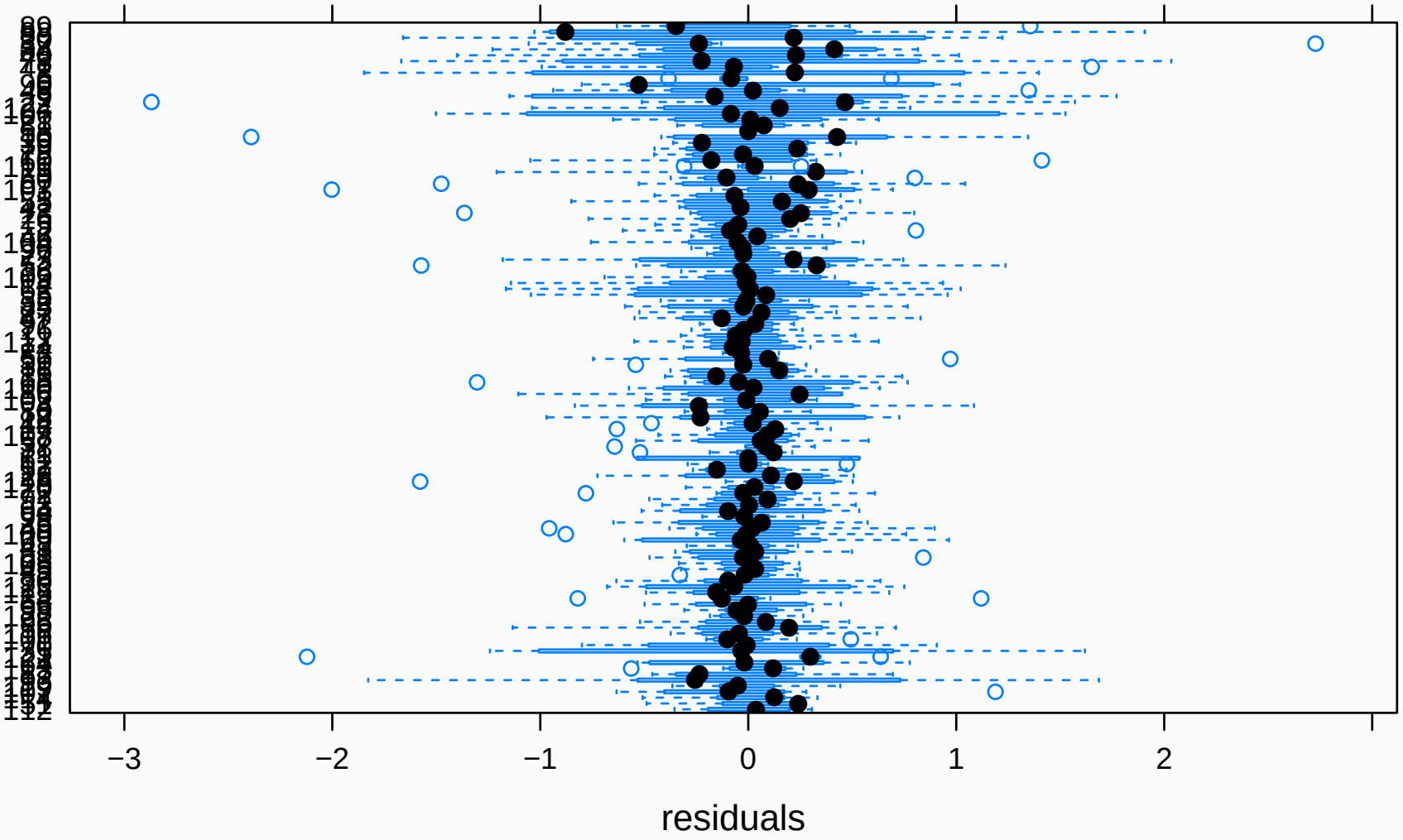
- Model 3: Regression lines with different intercept and slope for each class

$$\texttt{log(PP)}_{it} = \beta_{0i} + \beta_{1i} \, \texttt{YEAR}_{it} + \varepsilon_{it}$$

Residuals of Model 1 grouped by CLASS

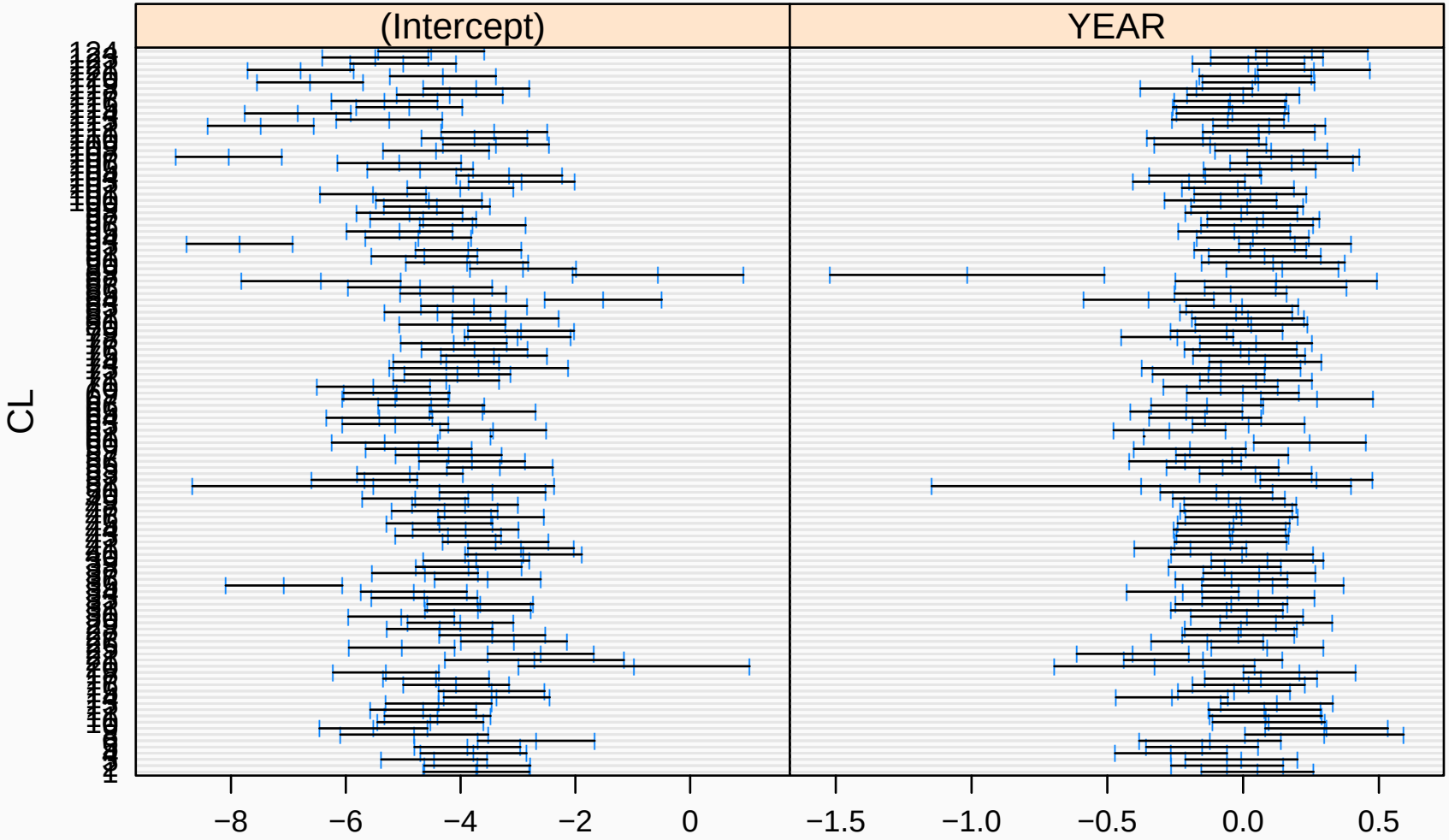


Residuals of Model 2 grouped by CLASS



Estimated intercepts and slopes for Model 3

Regression lines with different intercept and slope for each class



Workers Compensation: what the data say

Clear evidence of strong heterogeneity among classes, with some mild suggestion of separate intercepts and slopes (Model 3, with separate regression lines, has the smallest AIC, though the value is close to that of Model 2).

Modelling the pure premium with many separate regression lines is somewhat unsatisfactory, as we are fitting 118 separate models to small data sets of sample size 7 or less: we are not pooling information across classes.

Linear mixed models, which employ **random effects**, overcome such shortcomings, allowing for proper fitting to data and *borrowing information from all classes*.

Some theory

Longitudinal data models

The general basic structure of **linear mixed models**, in short **LMMs**, for longitudinal data is given by

$$y_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \boldsymbol{\alpha}_i + \varepsilon_{it}$$

with $i = 1, \ldots, n$ index for the subject (class, policy holder, or other entities), and $t = 1, \ldots, T_i \leq T$ index for the occasion.

$\mathbf{x}_{it} = (x_{0it}, \ldots, x_{(K-1)it})^\top$	$K \times 1$ vector of explanatory variables
$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{K-1})^\top$	$K \times 1$ vector of coefficients
$\mathbf{z}_{it} = (z_{0it}, \ldots, z_{(q-1)it})^\top$	$q \times 1$ vector of explanatory variables
$\boldsymbol{\alpha}_i = (\alpha_{0i}, \ldots, \alpha_{(q-1)i})^\top$	$q \times 1$ vector of **random coefficients**
ε_{it}	error term

Usual assumptions on the random components are

$\boldsymbol{\alpha}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ i.i.d. $\quad$ $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ i.i.d. $\quad$ $\boldsymbol{\alpha}_i$ and ε_{it} independent

Linear mixed models: comments

The word *mixed* refers to the presence of both fixed and random coefficients in the model.

The specification of a random distribution for the coefficients α_i has several advantages

- It models subject heterogeneity in a very flexible fashion.
- All the subjects are employed for estimating the fixed effects β and the variance $\mathbf{D}$, which characterizes the distribution of α_i.
- Both **prediction for a subject already in the study** or **prediction for a new subject** are possible. The latter is totally ruled out when the α_i are modelled as fixed effects.

The model requires that the available data come from a random sample of subjects.

Random intercepts model

A simple linear mixed model is obtained when $q = 1$ and the only random coefficients are just different intercepts for each subject

$$y_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}\,, \qquad \alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)\,\text{indep}\,.$$

The form for the **conditional** (on the random effects) mean response is

$$\mathrm{E}(Y_{it}|\alpha_i) = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \alpha_i\,,$$

whereas the conditional variance is

$$\mathrm{V}(Y_{it}|\alpha_i) = \sigma_\varepsilon^2\,.$$

The conditional covariance between two observations for the same subject is

$$\mathrm{Cov}(Y_{it}, Y_{it'}|\alpha_i) = 0\,,$$

observations are independent given the random effects.

Random intercepts model

A simple linear mixed model is obtained when $q = 1$ and the only random coefficients are just different intercepts for each subject

$$y_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}\,, \qquad \alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)\,\text{indep}\,.$$

The form for the **unconditional** mean response is

$$\mathrm{E}(Y_{it}) = \mathbf{x}_{it}^\top \boldsymbol{\beta}\,,$$

whereas the unconditional variance is

$$\mathrm{V}(Y_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2\,.$$

The unconditional covariance between two observations for the same subject is

$$\mathrm{Cov}(Y_{it}, Y_{it'}) = \sigma_\alpha^2\,,$$

so that their correlation is given by the **intraclass correlation** $\dfrac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$

One-way ANOVA model

The **one-way anova model** is just a random intercepts model without any covariate, namely $K = q = 1$, $x_{it} = z_{it} = 1$ and

$$y_{it} = \mu + \alpha_i + \varepsilon_{it} \, .$$

If y_{it} represents the response of the i-th subject in period t, then μ is the overall mean for all the subjects, and α_i the deviation of the hypothetical mean of the i-th subject from the overall mean μ.

There are three parameters that have to be estimated

$$\mu, \sigma^2_\alpha, \sigma^2_\varepsilon$$

It is also of interest to obtain plausible values for the random intercepts α_i. As these are random variables, it is customary to talk about **random effects prediction**, usually done by means of the **BLUP** method (on whose rationale we shall return later).

One-way ANOVA model: parameter estimates & prediction

The estimate of μ for fixed variances $\boldsymbol{\tau} = (\sigma_\alpha^2, \sigma_\varepsilon^2)$ is

$$\widehat{\mu}^{(\boldsymbol{\tau})} = \frac{\sum_{i=1}^{n} \zeta_i \, \overline{y}_i}{\sum_{i=1}^{n} \zeta_i} \,,$$

with $\overline{y}_i = \dfrac{1}{T_i} \sum_{t=1}^{T_i} y_{it}$ is the sample mean of the claims for the i-th subject, and $\zeta_i = \dfrac{T_i}{(T_i + \sigma_\varepsilon^2/\sigma_\alpha^2)}$.

The BLUP prediction of α_i is

$$\widehat{\alpha}_i^{(\boldsymbol{\tau})} = \zeta_i \, (\overline{y}_i - \widehat{\mu}^{(\boldsymbol{\tau})}) \,.$$

and the BLUP prediction of $y_{i,t+1}$ is

$$\widehat{y}_{i,t+1}^{(\boldsymbol{\tau})} = (1 - \zeta_i) \, \widehat{\mu}^{(\boldsymbol{\tau})} + \zeta_i \, \overline{y}_i \,,$$

Further simplifications occur in the **balanced case**, where $T_i = T$.

Shrinkage effect for one-way ANOVA model

The formula for the BLUP predictions for the one-way ANOVA model,

$$\widehat{y}_{i,t+1}^{(\boldsymbol{\tau})} = (1 - \zeta_i)\, \widehat{\mu}^{(\boldsymbol{\tau})} + \zeta_i\, \overline{y}_i\,,$$

highlights a general feature of random effect models.

The BLUP prediction of $y_{i,t+1}$ is actually a weighted mean of

- the subject sample mean $\overline{y}_i$, with weight ζ_i;
- the estimated overall mean $\widehat{\mu}^{(\boldsymbol{\tau})}$, with weight $1 - \zeta_i$.

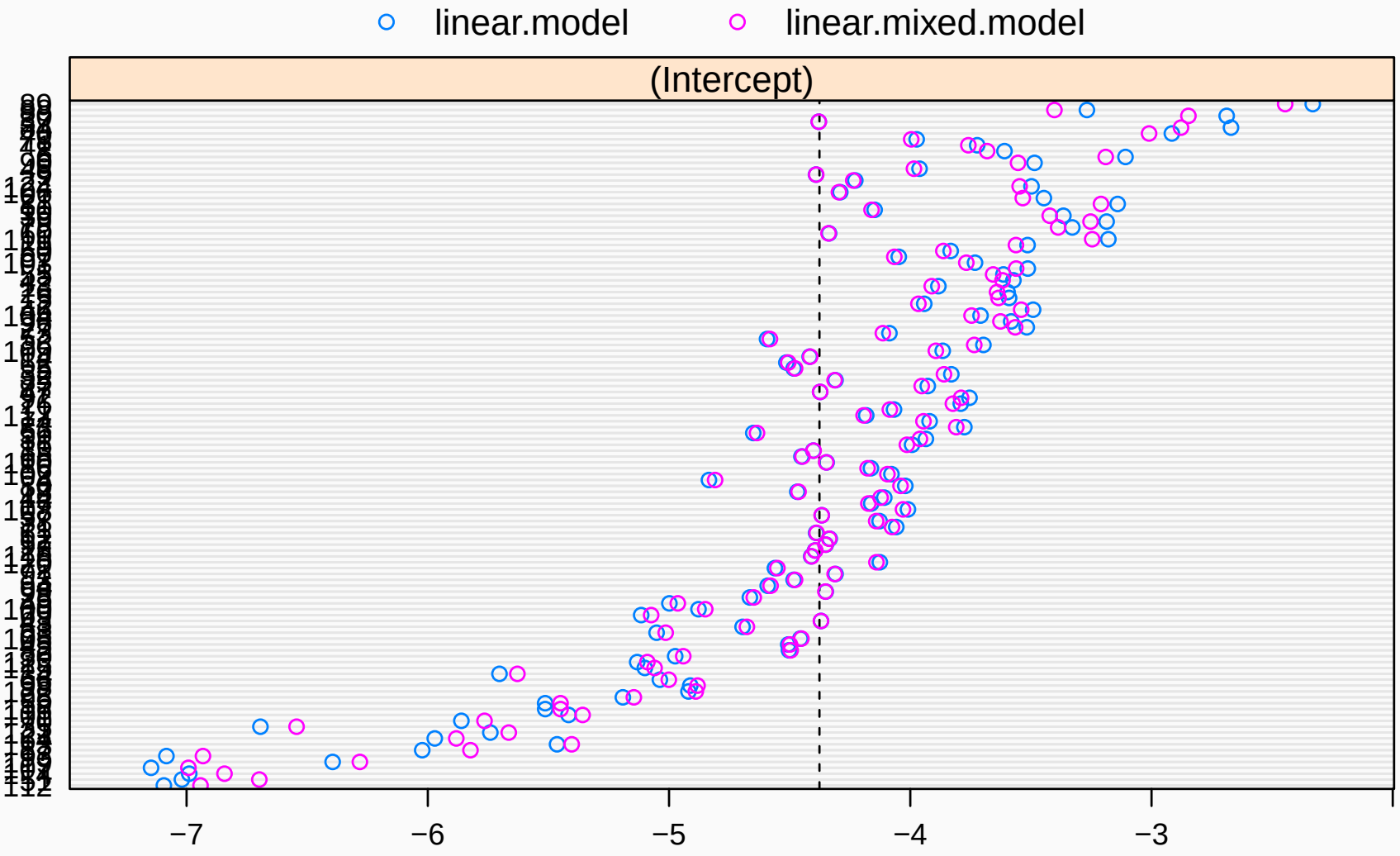
As $0 < \zeta_i < 1$, it follows that

$$|\widehat{y}_{i,t+1}^{(\boldsymbol{\tau})} - \widehat{\mu}^{(\boldsymbol{\tau})}| < |\overline{y}_i - \widehat{\mu}^{(\boldsymbol{\tau})}|$$

and this fact goes under the name of **shrinkage effect**.

The shrinkage effect explains why linear mixed models outperforms standard linear models for prediction!

Shrinkage effect: visualization



Random intercepts and slopes model

A random intercepts and slopes model fits separate regression lines for each subject.

The regression intercepts and slopes model is a special case with a subject-specific linear trend for time conditional on the random effects

$$y_{it} = (\beta_0 + \alpha_{0i}) + (\beta_1 + \alpha_{1i})t + \varepsilon_{it}\,,$$

with the usual assumption for the random effects

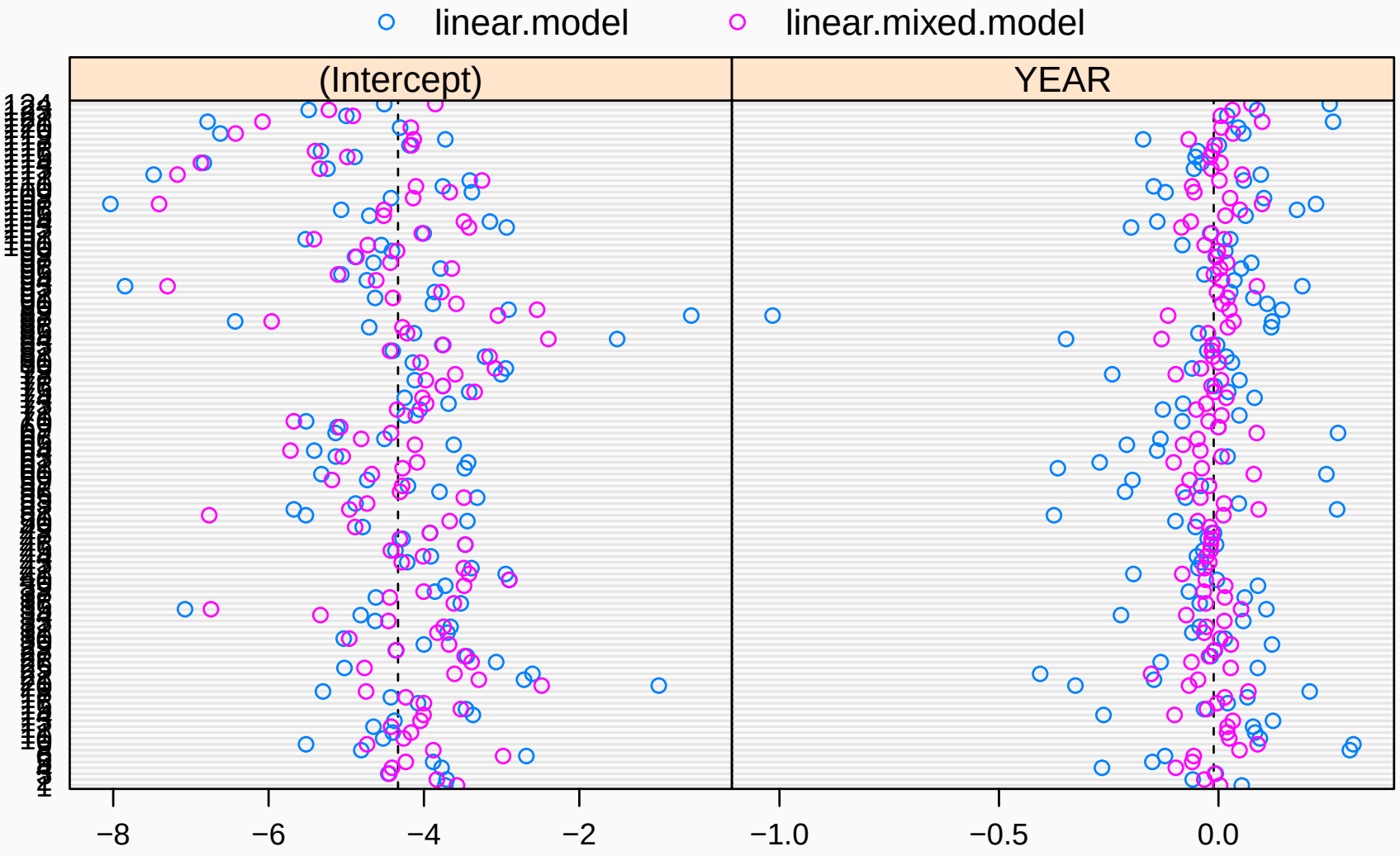
$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{0\boldsymbol{\alpha}} & \sigma_{01\boldsymbol{\alpha}} \\ \sigma_{01\boldsymbol{\alpha}} & \sigma^2_{1\boldsymbol{\alpha}} \end{pmatrix} \right\}.$$

The unconditional form for the mean response is a linear trend for time

$$\mathrm{E}(Y_{it}) = \beta_0 + \beta_1\, t\,,$$

and it is easy to verify that the variance changes with time.

Random intercepts and slopes for `log(PP)`: shrinkage effect



Matrix form of the model

Back to the general form of the model, it is useful to write the matrix form for the data of the i-th subject

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\alpha_i + \varepsilon_i$$

where

- $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})^\top$, $T_i \times 1$ vector of observations
- $\mathbf{X}_i$ is the $T_i \times K$ matrix of explanatory variables with t-th row $\mathbf{x}_{it}^\top$
- $\mathbf{Z}_i$ is the $T_i \times q$ matrix of explanatory variables with t-th row $\mathbf{z}_{it}^\top$
- $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT_i})^\top$, $T_i \times 1$ vector of error terms

Here $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$, with the $T_i \times T_i$ matrix $\mathbf{R}_i$ collecting all the parameters of the error term distribution.

Matrix form of the model

Stacking all the n subjects we obtain

$$\mathbf{y} = \mathbf{X}\,\beta + \mathbf{Z}\,\boldsymbol{\alpha} + \boldsymbol{\varepsilon}\,,$$

where all the vectors and matrices involved have $N = \sum_{i=1}^{n} T_i$ rows.

In compact form, conditional on the random effects

$$\begin{aligned} \mathrm{E}(\mathbf{Y}|\boldsymbol{\alpha}) &= \mathbf{X}\,\beta + \mathbf{Z}\,\boldsymbol{\alpha} \\ \mathrm{V}(\mathbf{Y}|\boldsymbol{\alpha}) &= \mathbf{R} \end{aligned}$$

Unconditionally, we get

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= \mathbf{X}\,\beta \\ \mathrm{V}(\mathbf{Y}) &= \mathbf{Z}\,\mathbf{D}\,\mathbf{Z}^{\top} + \mathbf{R} = \mathbf{V} \end{aligned}$$

Estimation of model parameters

Model parameters are given by the fixed effects β and by all the parameters needed to specify the distribution of both α and ε, here denoted by τ. The latter are often referred to as **variance components**.

If τ is known, β is estimated by the **Generalized Least Squares** (GLS) estimate

$$\widehat{\beta}^{(\tau)} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y}\,.$$

Usually, (β, τ) are estimated by maximum likelihood estimation, or variations of it, such as **restricted maximum likelihood estimation (REML)**.

The final estimate of β is always given by the GLS estimate with the matrix $\mathbf{V}$ evaluated at $\widehat{\tau}$.

General linear prediction problem

A crucial task in linear mixed models concerns the prediction of random quantities. The problem can be cast in terms of general theory for prediction.

In the **general linear prediction problem** we observe a random vector $\mathbf{y}$ with mean $\mathbf{X}\boldsymbol{\beta}$ and variance matrix $\mathbf{V}$, and the aim is to predict the value w of a random variable W with mean $\boldsymbol{\lambda}^\top \boldsymbol{\beta}$ and variance σ^2_w.

The best linear (in $\mathbf{y}$) predictor of w is

$$\widehat{w} = \mathrm{E}(W) + \mathrm{Cov}(W, \mathbf{Y})\, \mathbf{V}^{-1}\{\mathbf{y} - \mathrm{E}(\mathbf{Y})\}$$

which becomes

$$\widehat{w} = \boldsymbol{\lambda}^\top \boldsymbol{\beta} + \mathrm{Cov}(W, \mathbf{Y})\, \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\,.$$

Under normality, the formula corresponds to $\mathrm{E}(W|\mathbf{Y})$.

Best Linear Unbiased Prediction

In linear mixed models, the **Best Linear Unbiased Predictor** (BLUP) is obtained when the GLS estimator of β is inserted in the formula for $\widehat{w}$

$$\widehat{w}^{(\tau)} = \boldsymbol{\lambda}^\top \widehat{\beta}^{(\tau)} + \mathrm{Cov}(W, \mathbf{Y})\, \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\, \widehat{\beta}^{(\tau)})\,.$$

The formula assumes that $\mathrm{Cov}(W, \mathbf{Y})$ and $\mathbf{V}$ are known. Actual computations are carried out with estimated variance components.

Main applications of the BLUP formula are for $w = \mathbf{c}^\top \boldsymbol{\alpha}_i$ (for fixed $\mathbf{c}$), or for future observations, $w = y_{i,t+1}$: in either case, the key quantity is the BLUP predictor of $\boldsymbol{\alpha}_i$

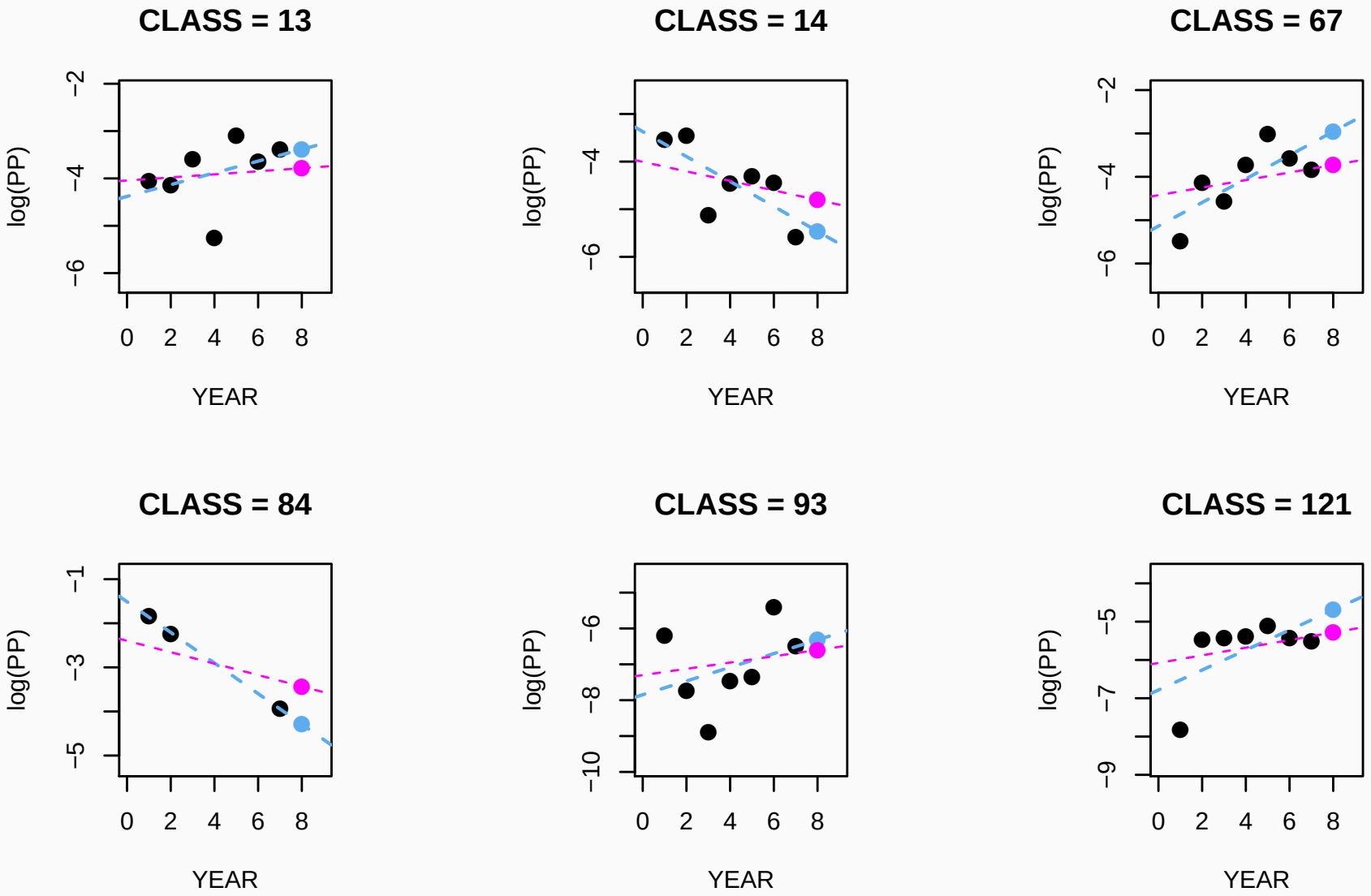
$$\widehat{\alpha}_i^{(\tau)} \quad = \quad \mathbf{D}\mathbf{Z}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i \widehat{\beta}^{(\tau)})\,,$$

which derives from the expression of $\mathrm{E}(\boldsymbol{\alpha}|\mathbf{Y})$.

Such formula is relevant with the name **empirical Bayes methods** often adopted for this kind of statistical techniques.

Random intercepts and slopes for `log(PP)`: shrinkage effect

Regression lines and predictions at `Year=8` for six selected classes:



Extensions

Bayesian methods for longitudinal data

Bayesian methods for linear mixed models add a prior distribution on the fixed effects β and the variance components τ.

The resulting models are just a special instance of **hierarchical Bayesian models**, where actually the difference between fixed and random effects is almost immaterial, being given by the different prior distributions employed for them.

At times prior information on both β and τ is vague, and a *diffuse prior* with large variance is assumed for them. In that case, inference on model parameters will be often similar to that provided by maximum likelihood.

MCMC methods have greatly facilitated the application of Bayesian methods for longitudinal data, with straightforward extension to prediction of random effects and future observations.

Longitudinal data with non-normal response

Longitudinal data and other hierarchical structures often involve a **non-normal response variable**. Notable examples are binary or count data.

In such cases the conditional distribution for the response $\mathbf{Y}$ given the random effects should be chosen in the GLM family.

The normality assumptions for the random effects, instead, is still a sensible one in most instances.

Generalized Linear Mixed Models

GLM specification for the conditional distribution of the response (given the random effects)

$+$

A set of normal random effects

$=$

Generalized Linear Mixed Models (GLMMs)

The interpretation of the various components of the model, and their usage in applications, are like in the linear case.

Generalized Additive Mixed Models

Nonlinear effects of continuous covariates can be inserted also in linear predictors of mixed models.

For example, a nonlinear effect of a certain covariate z in a random intercepts model may be modelled as

$$\eta_{it} = s(z_{it}) + \mathbf{x}_{it}^{\top} \beta + \alpha_i$$

The resulting models are called **Generalized Additive Mixed Models (GAMMs)**, and can be considered as an extension of both GAMs and GLMMs.

Like for GLMMs, computational issues are rather crucial for GAMMs, and in this sense such models are still in development. A full Bayesian approach may be convenient in many cases, though some care is required.