Mixed Models

Mixed models extend the predictor $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ of linear, generalized linear, and categorical regression models by incorporating random effects in addition to the nonrandom or "fixed" effects $\boldsymbol{\beta}$. Therefore, mixed models are sometimes also called random effects models and have become quite popular for analyzing longitudinal data obtained from repeated observations on individuals or objects in longitudinal studies. A closely related situation is the analysis of clustered data, i.e., when observations are obtained from objects selected by subsampling primary sampling units (clusters or groups of objects) in cross-sectional studies. For example, clusters may be defined by hospitals, schools, or firms, where data from (possibly small) subsamples of patients, students, or clients are collected. Generally, clustering may result from any data generating mechanism that induces a cluster structure. In any case, the data consist of n_i repeated observations

$$(y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ij}, \ldots, \boldsymbol{x}_{in_i})$$

of responses and covariates for each individual or cluster $i = 1, \ldots, m$.

For longitudinal data, y_{ij} and $\boldsymbol{x}_{ij}$ denote the observed value of the response and the covariate vector, respectively, for individual i at time t_{ij}, $j = 1, \ldots, n_i$, while for clustered data, these values are observations for subjects or objects j from cluster i. Mixed models allow estimation of individual- or cluster-specific effects, even in the case of relatively small numbers n_i of repeated individual measurements or sizes n_i of subsamples from clusters. The basic idea is to extend the linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ for observation y_{ij} with fixed population effects $\boldsymbol{\beta}$ to the linear mixed predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i$. Usually, $\boldsymbol{u}'_{ij}$ is a subvector of the covariates, and $\boldsymbol{\gamma}_i$ is a vector of individual- or cluster-specific random effects. The assumption of $\boldsymbol{\gamma}_i$ to be fixed effects as in the standard linear or generalized linear model is often impractical since the number of parameters to be estimated becomes quite large relative to the sample size. On the other hand, the random effects distribution implicitly induces certain regularization properties for the cluster parameters γ_i. An additional advantage of mixed models is that correlations, induced by repeated observations from individuals or clusters, are taken into account during estimation.

In the following, we first describe linear mixed models (LMMs) with (conditionally) Gaussian responses y_{ij}, making the conventional assumption that the random effects are i.i.d. Gaussian variables. We then extend LMMs by allowing correlated Gaussian random effects. This leads to a very broad class of models that are appropriate for analyzing spatial and spatio-temporal data, as well as for Bayesian approaches to non- and semiparametric regression in Chaps. 8 and 9, in particular in Sects. 8.1.9 and 9.6. Statistical inference is described from a frequentist likelihood-oriented as well as a Bayesian perspective.

The second part of the chapter extends generalized linear models for non-Gaussian responses to generalized linear mixed models (GLMMs), such as logit or Poisson regression models. Statistical inference for the GLMM is based on similar, but more complicated, concepts of the LMM.

7.1 Linear Mixed Models for Longitudinal and Clustered Data

7.1.1 Random Intercept Models

We start with random intercept models which are among the most simple (albeit quite important) mixed models. For notational simplicity, we first restrict ourselves to the case of just one covariate x. Let

$$(y_{ij}, x_{ij}), \quad i = 1, \dots, m, \quad j = 1, \dots, n_i$$

denote the values of the response variable y and covariate x observed at times $t_{i1} < \dots < t_{ij} < \dots < t_{in_i}$ for individuals $i = 1, \dots, m$ in a longitudinal study or for subjects $j = 1, \dots, n_i$ in clusters $i = 1, \dots, m$. Our starting point for modeling the relationship between y and x is the classical linear model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \tag{7.1}$$

with i.i.d. errors $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. In this model, the fact that we have *repeated measurements* $j = 1, \dots, n_i$ on the *same individual or cluster* i is not taken into account. In particular, we not only assume that the observations y_{ij} and y_{rl}, of different individuals i and r, are (stochastically) independent but also repeated measurements y_{ij} and y_{il} on the same individual or cluster i. A possible graphical way to check this independence assumption is to estimate and plot separate regression lines (or more generally regression curves) for each individual or cluster i. If there is no cluster-specific heterogeneity, all regression lines should have similar (not identical due to sampling variability) intercepts and slopes. A typical plot is shown in the left panel of Fig. 7.1. The estimates are based on artificial data drawn from the model $y_{ij} = 1 + x_{ij} + \varepsilon_{ij}$,

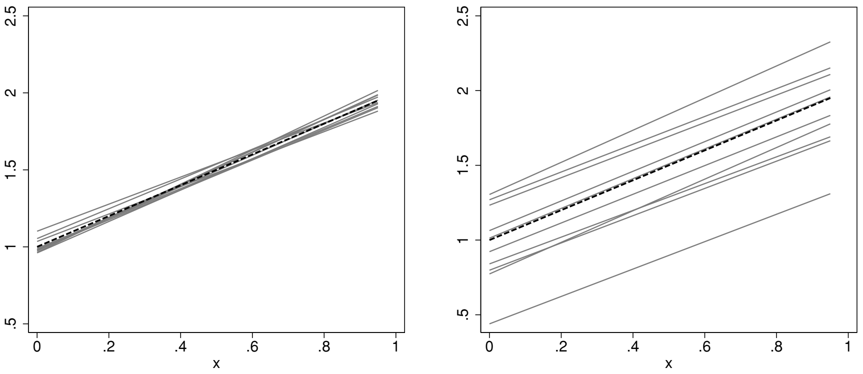

Fig. 7.1 Illustration of random intercept models: both panels show separately estimated regression lines for each cluster. In the *left panel*, there is no cluster-specific random intercept, while in the *right panel*, a random intercept is present. The *dashed line* corresponds to the population model

$i = 1, \ldots, 10$, $j = 1, \ldots, 20$, for $m = 10$ individuals or clusters, $n_i = 20$ repeated measurements in each cluster, and i.i.d. errors $\varepsilon_{ij} \sim \mathrm{N}(0, 0.1^2)$. Clearly, the estimated cluster-specific regression lines scatter with low variability around the true regression line $1 + x$ (dashed line in Fig. 7.1). Hence, there is no reason for assuming cluster-specific heterogeneity, and a common regression model for all clusters is sufficient.

The right panel of Fig. 7.1 reveals a different scenario. The estimated cluster-specific regression lines still show a common slope across clusters, but the intercept appears to be different from cluster to cluster. To model this type of cluster-specific heterogeneity, we introduce cluster-specific parameters γ_{0i} and obtain

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \varepsilon_{ij}, \tag{7.2}$$

where $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ are the usual i.i.d. errors of the classical linear model. In Eq. (7.2),

- β_0 is the "fixed" population intercept.
- γ_{0i} is the individual- or cluster-specific (random) deviation from the population intercept β_0.
- $\beta_0 + \gamma_{0i}$ is the (random) intercept for cluster i.
- β_1 is a "fixed" population slope parameter of covariate x that is common across clusters.

Since the individuals or clusters are a random sample from a larger population, the cluster-specific parameters γ_{0i} are assumed to be random with

$$\gamma_{0i} \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \tau_0^2\right). \tag{7.3}$$

We also assume mutual independence between the ε_{ij} and the γ_{0i}. The normal *random effects distribution* in Eq. (7.3) is also sometimes called a *mixture distribution*.

The mean in (7.3) can be set to zero because the population mean is already represented by the fixed effect β_0.

The random intercepts $\beta_0 + \gamma_{0i} \sim \mathrm{N}(\beta_0, \tau_0^2)$ may be interpreted as effects of omitted (individual- or cluster-specific) covariates and account for unobserved heterogeneity. Another way to look at the model is to interpret γ_{0i} as an additional error term. The random intercept model appears as a linear regression model with two error terms, where γ_{0i} is then a cluster-level error shared between measurements on the same individual or cluster i and ε_{ij} is the observation error of measurement j in cluster i.

The random intercept model induces a specific correlation or dependence structure on the responses y_{ij}. Given the random intercepts γ_{0i}, the y_{ij} are still conditionally independent with

$$y_{ij} \mid \gamma_{0i} \sim \mathrm{N}(\beta_0 + \beta_1 x_{ij} + \gamma_{0i}, \sigma^2).$$

Marginally, however, repeated measurements y_{ij} for subject or cluster i are correlated with within-subject correlation coefficient

$$\mathrm{Corr}(y_{ij}, y_{il}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}, \quad j \neq l; \tag{7.4}$$

see Sect. 7.1.4 for a derivation. Based on the normality assumption of the random effects and the errors, we can further derive the marginal distribution of responses. We have

$$\boldsymbol{y}_i \sim \mathrm{N}(\boldsymbol{X}_i \boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n_i} + \tau_0^2 \boldsymbol{J}_{n_i}), \tag{7.5}$$

where $\boldsymbol{X}_i$ is an $(n_i \times 2)$-design matrix with ones in the first column and the observed x_{ij} in the second column, and $\boldsymbol{J}_{n_i}$ denotes an $(n_i \times n_i)$-matrix of ones. Between two subjects i and r, the observations y_{ij} and y_{rl} are still uncorrelated. The strength of the within-subject correlation (7.4) depends on the magnitude of the error variances τ_0^2 and σ^2. The higher the random effects variance τ_0^2 relative to the error variance σ^2, the stronger the within-subject correlation. Note also that the within-subject correlation is constant (equicorrelation) from measurement to measurement. This may be questionable for longitudinal data, where we might expect correlations dying off for measurements which are farther apart in time.

The marginal model (7.5) also shows what happens with the estimates for $\boldsymbol{\beta}$ if the classical linear model (7.1) is estimated instead of the random intercept model (7.2). If the assumed correlation structure induced by the random intercept model is correct, estimating a classical linear model means that we mistakenly assume an error covariance matrix $\sigma^2 \boldsymbol{I}$ instead of a non-diagonal covariance matrix as imposed in the marginal model (7.5) of the random intercept model. The consequences of using an incorrect covariance matrix have already been established in Sect. 4.1.1 in the context of the general linear model. As stated there, the estimates for the "fixed" regression coefficients $\boldsymbol{\beta}$ are still unbiased. However, the covariance matrix of $\boldsymbol{\beta}$ and all derived quantities in particular standard errors, confidence intervals, and tests are not correct. Note that the standard errors could either be smaller or larger, as in a misspecified classical linear model. A nice description of the consequences of mistakenly specifying a classical linear model is also given in Skrondal and Rabe-Hesketh (2008) in Sect. 3.10.1.

Between- and Within-Cluster Effects

The random intercept model (7.2) considered thus far has an important limitation: The so-called within- and between-cluster effects of x are the same. The within-cluster effect denotes the effect if x changes within the *same individual* at different occasions. The between-cluster effect refers to different x values *between different subjects or clusters*. In either case, a difference of one unit of x in model (7.2) induces a difference of β_1 in expected responses. This equality between the within- and between-cluster effect might be questionable in applications. A possible example is our data on malnutrition in Zambia. Recall that the individual data on children are nested within the districts of Zambia as the cluster variable. Suppose we are interested in the effect of the households wealth, measured by a wealth index, x say, on the Z-score. Now we focus our attention on two imaginary districts, one comparably rich district i and a second rather poor district l. More precisely, we assume that in district i, the population is on average richer than in the second district l, i.e., $\bar{x}_i > \bar{x}_l$ with $\bar{x}_i, \bar{x}_l$ being the cluster averages of the wealth index. Suppose first that we compare a child living in the rich district with a child living in the poor district (between district comparison). It is then conceivable that the child living in the rich environment has a higher Z-score, i.e., is better nourished, than the child in the poor environment, even if the individual household wealth is identical. The reason might be that the child profits from the rich environment independent of the individual household situation that might be much less favorable. Economists call this an *external effect*. A possible way to model this between-cluster effect is simply to include the cluster averages $\bar{x}_i$ as an additional covariate into the regression equation. On the other hand, there might be also a within-cluster effect if the individual household wealth is different from that of the average cluster wealth. Children living in households which are wealthier than the average households in the district should have an even higher Z-score (at least on average). This effect may be equal in size to the between-cluster effect but could just as well differ, i.e., being smaller or larger in size. An illustration of different within- and between-cluster effects is shown in Fig. 7.2. The left panel shows different sizes of the within- and between-cluster effects but with identical signs. In the right panel, the signs of the within- and between-cluster effects are opposite of each other.

To deal with possibly different within- and between-cluster effects, we can incorporate two covariates derived from x into the predictor: The between-cluster effect can be modeled by the respective cluster means $\bar{x}_i$ as a covariate. The within-cluster effect is modeled by incorporating the individual difference from the cluster mean $x_{ij} - \bar{x}_i$. This yields the extended random intercept model

$$y_{ij} = \beta_0 + \beta_1(x_{ij} - \bar{x}_i) + \beta_2\bar{x}_i + \gamma_{0i} + \varepsilon_{ij}, \tag{7.6}$$

where the coefficient β_2 of the cluster mean represents the between-cluster effect, and the coefficient β_1 of the individual deviation from the cluster mean repre-

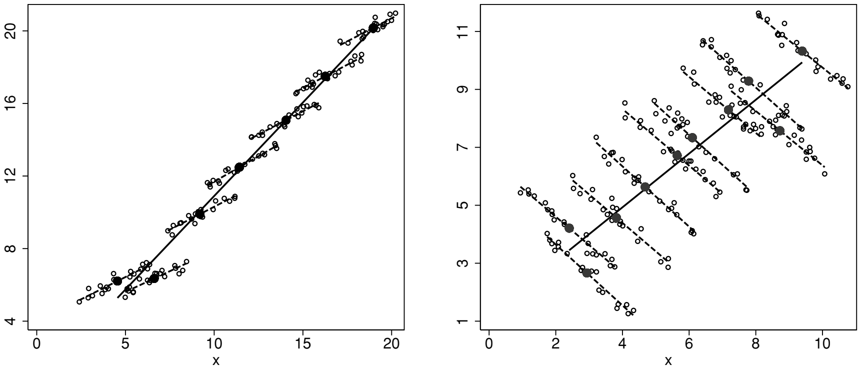

Fig. 7.2 Illustration of within- and between-cluster effects: the between-cluster effects are visualized through the cluster means $\bar{x}_i$ marked by *black dots* and the corresponding linear trends (*solid lines*) which are increasing with x in both panels. The within-cluster effects are illustrated by *dashed lines*

sents the within-cluster effect. The model collapses to the original random intercept model (7.2) if the within- and between-cluster effects are identical, i.e., if $\beta_1 = \beta_2$. Another interpretation of model (7.6) is obtained by considering $\tilde{\gamma}_{0i} = \beta_2 \bar{x}_i + \gamma_{0i}$ as a random intercept that depends through $\beta_2 \bar{x}_i$ on the covariates.

Alternative Views on the Random Intercept Model

For some applications, the view of the γ_{0i} as random effects can be questionable. This is the case when we do not have the interpretation that clusters are randomly sampled from a larger population. Such a situation arises, for example, when data have been observed on a discrete spatial grid. A typical example is the data for the Munich rent index where the district of each apartment in Munich is given. To account for spatial heterogeneity, it might be useful to add a district-specific effect into the predictor; see Example 9.2 of Chap. 9. Since the districts cannot be seen as a random sample of a larger "population" of districts, the interpretation of the district-specific effects as random effects is somewhat artificial. However, there are alternative useful interpretations of the random intercept model which are more suitable for the given situation. In particular, the random effects distribution (7.3) can be readily understood as the prior for γ_{0i} in a corresponding Bayesian approach. In fact, Eq. (7.3) is identical to the Bayesian ridge prior of Sect. 4.4.2.1. Assuming noninformative priors for the "fixed" effects, i.e., $p(\beta_0) \propto \text{const}$ and $p(\beta_1) \propto \text{const}$, the posterior is given by

$$p(\beta_0, \beta_1, \boldsymbol{\gamma} \mid \boldsymbol{y}) \propto L(\beta_0, \beta_1, \boldsymbol{\gamma}) \prod_{i=1}^{m} \frac{1}{\sqrt{\tau_0^2}} \exp\left(-\frac{1}{2\tau_0^2}\gamma_{0i}^2\right),$$

where $\boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0m})'$ is the vector of cluster effects and $L(\cdot)$ is the Gaussian likelihood of the conditional model. The variances σ^2 and τ_0^2 are assumed fixed for the moment. Now the posterior mode can be obtained by maximizing the log-posterior resulting in the optimization criterion

$$\mathrm{PLS}(\beta_0, \beta_1, \gamma) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \beta_0 - \beta_1 x_{ij} - \gamma_{0i})^2 + \lambda \sum_{i=1}^{m} \gamma_{0i}^2 \tag{7.7}$$

with $\lambda = \sigma^2/\tau_0^2$. This has the form of a penalized least squares criterion quite similar to ridge regression outlined in Sect. 4.2.2. To understand the nature of the penalization, we consider the particularly simple random intercept model

$$y_{ij} = \beta_0 + \gamma_{0i} + \varepsilon_{ij}, \tag{7.8}$$

without any covariates, and as usual with $\gamma_{0i} \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau_0^2)$, $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$. As will be shown in Sect. 7.3.2, the estimator for the γ_{0i} is given by

$$\hat{\gamma}_{0i} = \frac{n_i \tau_0^2}{\sigma^2 + n_i \tau_0^2} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta}_0),$$

where $\hat{\beta}_0 = \bar{y}$, with $\bar{y}$ the overall mean of the responses. The estimator for the cluster mean $\eta_i = \beta_0 + \gamma_{0i}$ is now given by

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\gamma}_{0i} = \bar{y} + \frac{n_i \tau_0^2}{\sigma^2 + n_i \tau_0^2} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}). \tag{7.9}$$

The term $e_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})$ can be seen as an average residual for individual i, which is a rather natural estimate of γ_{0i}. This is multiplied by the factor

$$\lambda_i = \frac{n_i \tau_0^2}{\sigma^2 + n_i \tau_0^2} < 1,$$

which is sometimes called a shrinkage effect, because the ad hoc estimate e_i for γ_i is shrunken toward the prior mean 0. The larger the n_i, the closer the weight λ_i is to 1 and the smaller the shrinkage. Additional shrinkage is obtained if the error variance σ^2 is large relative to the random effects variance τ_0^2.

It is instructive to contrast the estimator (7.9) with two extreme modeling strategies:

- *Full ignorance of groups:* On the one extreme, we could fully ignore the groups and estimate the model $y_{ij} = \beta_0 + \varepsilon_{ij}$ with fixed overall intercept β_0. This is also called the fully pooled model. Of course, the least squares estimator for β_0 (which is then identical to the cluster mean η_i) is given by the overall mean $\bar{y}$ of responses, i.e., $\hat{\eta}_i = \hat{\beta}_0 = \bar{y}$.

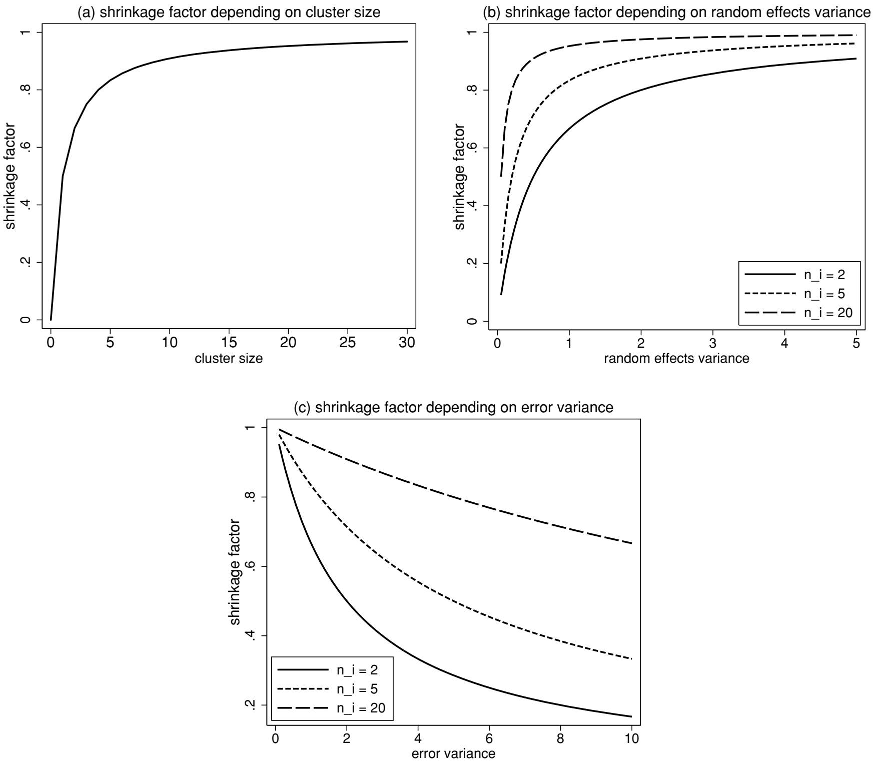

Fig. 7.3 Shrinkage factor in simple random intercept models: Panel **a** illustrates the shrinkage factor λ_i depending on the cluster size n_i (for fixed $\sigma^2 = 1$ and $\tau_0^2 = 1$). Panels **b** and **c** display the shrinkage factor depending on the random effects variance τ_0^2 (for fixed $\sigma^2 = 1$) and the error variance σ^2 (for fixed $\tau_0^2 = 1$), respectively, and three different choices for the cluster size n_i

- *Full distinction of groups:* The other extreme would be to estimate separate models $y_{ij} = \eta_i + \varepsilon_{ij}$ for each cluster i treating the η_i as fixed parameters (without random effects distributions) in a model without an intercept (to omit collinearity problems). This is also known as fully unpooled estimation. Here, the least squares estimators for the η_i are given by the cluster means $\bar{y}_i$, i.e., $\hat{\eta}_i = \bar{y}_i$.

Now the random effects estimator (7.9) can be seen as a compromise between the two extreme cases. For large n_i or small σ^2 relative to τ_0^2, the estimator (7.9) approaches the fully unpooled estimator $\bar{y}_i$. For small n_i or large σ^2 relative to τ_0^2, the fully pooled estimator $\bar{y}$ is approached. In the extreme cases $n_i = 0$ or $\tau_0^2 \to 0$ or $\sigma^2 \to \infty$, the mean $\bar{y}$ is reached as a limit. In the other extreme cases $n_i \to \infty$ or $\tau_0^2 \to \infty$ or $\sigma^2 \to 0$, we reach the cluster mean $\bar{y}_i$ as a limit. An illustration of the shrinkage factor λ_i depending on the cluster size n_i (panel a), the random effects variance τ_0^2 (panel b), and the error variance σ^2 (panel c) can be found in Fig. 7.3.

Key Features of Mixed Models

Although the random intercept model is comparably simple, it already reveals the key features and advantages of mixed models:

- Individual- or cluster-specific effects can be introduced to account for specific deviations from the population behavior.
- They allow to correct for unobserved heterogeneity induced by omitted covariates.
- Correlations between observations of the same individual or cluster can be taken into account (at least to some extent). This ensures that inference regarding the regression coefficients is correct in the sense that we obtain correct standard errors, confidence intervals, and tests.
- Estimation is stabilized by assuming a common random effects distribution that acts as a penalty term for the otherwise unpenalized cluster-specific effects.

We finally summarize the various interpretations of the random intercept model:

- A *classical interpretation*, where clusters are a random sample of a larger population and the γ_{0i} are cluster-specific random effects.
- A *marginal interpretation*, where the random effects γ_{0i} induce the general linear model (7.5) with correlated errors for the observed y_{ij}.
- A *Bayesian point of view*, which interprets Eq. (7.3) as an underlying prior.
- A *penalized least squares view*, where the penalized least squares criterion induces a penalty on the cluster-specific effects to regularize the estimated parameters, similar as in ridge regression; see Sect. 4.2.2.