3

Diagnostic Checks for Multilevel Models

Tom A. B. Snijders[1,2] and Johannes Berkhof[3]

[1] University of Oxford
[2] University of Groningen
[3] VU University Medical Center, Amsterdam

3.1 Specification of the Two-Level Model

This chapter focuses on diagnostics for the two-level Hierarchical Linear Model (HLM). This model, as defined in chapter 1, is given by

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j\,, \quad j = 1, \ldots, m, \tag{3.1a}$$

with

$$\begin{pmatrix} \underline{\boldsymbol{\epsilon}}_j \\ \underline{\boldsymbol{\delta}}_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\emptyset} \\ \boldsymbol{\emptyset} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_j(\boldsymbol{\theta}) & \boldsymbol{\emptyset} \\ \boldsymbol{\emptyset} & \boldsymbol{\Omega}(\boldsymbol{\xi}) \end{pmatrix} \right) \tag{3.1b}$$

and

$$(\underline{\boldsymbol{\epsilon}}_j, \underline{\boldsymbol{\delta}}_j) \perp (\underline{\boldsymbol{\epsilon}}_\ell, \underline{\boldsymbol{\delta}}_\ell) \tag{3.1c}$$

for all $j \neq \ell$. The lengths of the vectors $\boldsymbol{y}_j$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}_j$, respectively, are n_j, r, and s. Like in all regression-type models, the explanatory variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are regarded as fixed variables, which can also be expressed by saying that the distributions of the random variables $\underline{\boldsymbol{\epsilon}}$ and $\underline{\boldsymbol{\delta}}$ are conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$. The random variables $\underline{\boldsymbol{\epsilon}}$ and $\underline{\boldsymbol{\delta}}$ are also called the vectors of residuals at levels 1 and 2, respectively. The variables $\underline{\boldsymbol{\delta}}$ are also called random slopes. Level-two units are also called clusters.

The standard and most frequently used specification of the covariance matrices is that level-one residuals are i.i.d., i.e.,

$$\boldsymbol{\Sigma}_j(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_{n_j}\,, \tag{3.1d}$$

where $\boldsymbol{I}_{n_j}$ is the n_j-dimensional identity matrix; and that either all elements of the level-two covariance matrix $\boldsymbol{\Omega}$ are free parameters (so one could identify $\boldsymbol{\Omega}$ with $\boldsymbol{\xi}$), or some of them are constrained to 0 and the others are free parameters.

Handbook of Multilevel Analysis, edited by Jan de Leeuw and Erik Meijer

Questioning this model specification can be aimed at various aspects: the choice of variables included in $\boldsymbol{X}$, the choice of variables for $\boldsymbol{Z}$, the residuals having expected value 0, the homogeneity of the covariance matrices across clusters, the specification of the covariance matrices, and the multivariate normal distributions. Note that in our treatment the explanatory variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ are regarded as being deterministic; the assumption that the expected values of the residuals (for fixed explanatory variables!) are zero is analogous to the assumption, in a model with random explanatory variables, that the residuals are uncorrelated with the explanatory variables.

The various different aspects of the model specification are entwined, however: problems with one may be solved by tinkering with one of the other aspects, and model misspecification in one respect may lead to consequences in other respects. E.g., unrecognized level-one heteroscedasticity may lead to fitting a model with a significant random slope variance, which then disappears if the heteroscedasticity is taken into account; non-linear effects of some variables in $\boldsymbol{X}$, when unrecognized, may show up as heteroscedasticity at level one or as a random slope; and non-zero expected residuals sometimes can be dealt with by transformations of variables in $\boldsymbol{X}$.

This presentation of diagnostic techniques starts with techniques that can be represented as model checks remaining within the framework of the HLM. This is followed by a section on model checking based on various types of residuals. An important type of misspecification can reside in non-linearity of the effects of explanatory variables. The last part of the chapter presents methods to identify such misspecifications and estimate the non-linear relationships that may obtain.

3.2 Model Checks within the Framework of the Hierarchical Linear Model

The HLM is itself already a quite general model, a generalization of the General Linear Model, the latter often being used as a point of departure in modeling or conceptualizing effects of explanatory on dependent variables. Accordingly, checking and improving the specification of a multilevel model in many cases can be carried out while staying within the framework of the multilevel model. This holds to a much smaller extent for the General Linear Model. This section treats some examples of model specification checks which do not have direct parallels in the General Linear Model.

3.2.1 Heteroscedasticity

The comprehensive nature of most algorithms for estimating the HLM makes it relatively straightforward to include some possibilities for modeling het-

eroscedasticity, i.e., non-constant variances of the random effects. (This is sometimes indicated by the term "complex variation", which however does not imply any thought of the imaginary number $i = \sqrt{-1}$.)

As an example, the iterated generalized least squares (IGLS) algorithm implemented in MLwiN [18, 19] accommodates variances depending as linear or quadratic functions of variables. For level-one heteroscedasticity, this is carried out formally by writing

$$\underline{\epsilon}_{ij} = \boldsymbol{v}_{ij}\underline{\boldsymbol{\epsilon}}^0_{ij}$$

where $\boldsymbol{v}_{ij}$ is a $1 \times t$ variable and $\underline{\boldsymbol{\epsilon}}^0_{ij}$ is a $t \times 1$ random vector with

$$\underline{\boldsymbol{\epsilon}}^0_{ij} \sim \mathcal{N}\big(\boldsymbol{\emptyset}, \boldsymbol{\Sigma}^0(\boldsymbol{\theta})\big).$$

This implies

$$\mathrm{Var}(\underline{\epsilon}_{ij}) = \boldsymbol{v}_{ij}\boldsymbol{\Sigma}^0(\boldsymbol{\theta})\boldsymbol{v}'_{ij}\,. \tag{3.3}$$

The standard homoscedastic specification is obtained by letting $t = 1$ and $\boldsymbol{v}_{ij} \equiv 1$.

The IGLS algorithm works only with the expected values and covariance matrices of $\underline{\boldsymbol{y}}_j$ implied by the model specification, see Goldstein [18, pp. 49–51]. A sufficient condition for model (3.1a)–(3.1c) to be a meaningful representation is that (3.3) is nonnegative for all i, j — clearly less restrictive than $\boldsymbol{\Sigma}^0$ being positive definite. Therefore it is not required that $\boldsymbol{\Sigma}^0$ be positive definite, but it is sufficient that (3.3) is positive for all observed $\boldsymbol{v}_{ij}$. E.g., a level-one variance function depending linearly on $\boldsymbol{v}$ is obtained by defining

$$\boldsymbol{\Sigma}^0(\boldsymbol{\theta}) = \big(\sigma_{hk}(\boldsymbol{\theta})\big)_{1\le h,k\le t}$$

with

$$\begin{aligned} &\sigma_{h1}(\boldsymbol{\theta}) = \sigma_{1h}(\boldsymbol{\theta}) = \theta_h, && h = 1, \ldots, t \\ &\sigma_{hk}(\boldsymbol{\theta}) = 0, && \min\{h, k\} \ge 2 \end{aligned}$$

where $\boldsymbol{\theta}$ is a $t \times 1$ vector. Quadratic variance functions can be represented by letting $\boldsymbol{\Sigma}^0$ be a symmetric matrix, subject only to a positivity restriction for (3.3).

In exactly the same way, variance functions for the level-two random effects depending linearly or quadratically on level-two variables are obtained by including these level-two variables in the matrix $\boldsymbol{Z}$. The usual interpretation of a "random slope" then is lost, although this term continues to be used in this type of model specification.

Given that among multilevel modelers random slopes tend to be more popular than heteroscedasticity, unrecognized heteroscedasticity may show up in the form of a fitted model with a random slope of the same or a correlated variable, which then may disappear if the heteroscedasticity is modeled.

Therefore, when a researcher is interested in a random slope of some variable Z_k and thinks to have found a significant slope variance, it is advisable to test for the following two kinds of heteroscedasticity: the level-one residual variance may depend (e.g., linearly or quadratically) on the variable Z_k, or the level-two intercept variance may depend on the cluster mean of Z_k, i.e., on the variable defined by

$$\bar{z}_{.jk} = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{ijk} \ .$$

Given that one uses software that can implement models with these types of heteroscedasticity, this is an easy (and sometimes disconcerting) model check. Some examples of checking for heteroscedasticity can be found in Goldstein [18, Chapter 3] and Snijders and Bosker [51, Chapter 8].

3.2.2 Random or Fixed Coefficients

A basic question in applying the HLM is whether a random coefficient model is appropriate at all for representing the differences between the level-two units. In other words, is it appropriate indeed to treat the variables $\underline{\boldsymbol{\delta}}_j$ in (3.1) as random variables, or should they rather be treated as fixed parameters $\boldsymbol{\delta}_j$?

On a conceptual level, this depends on the purpose of the statistical inference. If the level-two units j may be regarded as a sample from some population (which in some cases will be hypothetical or hard to circumscribe, but nevertheless conceptually meaningful) and the statistical inference is directed at this population, then a random coefficient model is in principle appropriate; cf. Hsiao [30]. This is the case, e.g., when one wishes to test the effect of an explanatory variable that is defined at level two, i.e., it is a function of the level-two units only. Then testing this variable has to be based on some way of comparing the variation accounted for by this variable to the total residual variation between level-two units, and it is hard to see how this could be done meaningfully without assuming that the level-two units are a sample from a population.

If, on the other hand, the statistical inference aims only at the particular set of units j included in the data set at hand, then a fixed effects model is appropriate. Note that in the fixed effects model the only random effects are the level-one residuals $\underline{\boldsymbol{\epsilon}}_j$; under the usual assumption (3.1d) of homoscedasticity, this model can be analysed by ordinary least squares (OLS) regression, so that the analysis is very straightforward except perhaps for the large number of dummy variables. When the cluster sizes are very large, there is hardly a difference between the fixed effects and the random effects specification for the estimation of parameters that they have in common.

If the differences between the level-two units are a nuisance factor rather than a point of independent interest, so that there is interest only in the

within-cluster effects, the analysis could in principle be done either way. Then the fixed effects estimates of the within-cluster regression coefficients, obtainable by OLS regression, achieve a better control for unexplained differences between the level-two units, because they do not need the assumption that the explanatory variables $\boldsymbol{X}$ are uncorrelated with the level-two random effects $\underline{\boldsymbol{\delta}}$. More generally, the fixed effects estimates have the attractive robustness property that they are not influenced at all by the specification of the level-two model. This can of course be generalized to models with more than two levels. This robustness property is elaborated with a lot of detailed matrix calculus in Kim and Frees [32].

On a practical level, the choice between random and fixed effects depends strongly on the tenability of the model assumptions made for the random coefficients and the properties of the statistical procedures available under the two approaches. Such practical considerations will be especially important if the differences between level-two units are a nuisance factor only. The assumptions in model (3.1) for the random effects $\underline{\boldsymbol{\delta}}_j$ are their zero expectations, homogeneous variances, and normal distributions. The normality of the distributions can be checked to some extent by plots of residuals (see below). If normality seems untenable, one could use models with other distributions for the random effects such as t-distributions (e.g., Seltzer, Wong, and Bryk [48]) or mixtures of normal distributions (the heterogeneity model of Verbeke and Lesaffre [54], also see Verbeke and Molenberghs [55]). Homogeneity of the variances is very close to the assumption that the level-two units are indeed a random sample from a population; in the preceding section it was discussed how to model variances depending on level-two variables, which can occur, e.g., if the level-two units are a sample from a stratified population and the variances depend on the stratum-defining variables.

To understand the requirement that the expected values of the level-two residuals are zero, we first focus on the simplest case of a random intercept model, where $\boldsymbol{Z}_j$ contains only the constant vector with all its n_j entries equal to 1, expressed as $\boldsymbol{Z}_j = \mathbf{1}_{n_j}$. Subsequently we shall give a more formal treatment of a more general case.

The level-two random effects $\underline{\boldsymbol{\delta}}_j$ consist of only one variable, the random intercept $\underline{\delta}_j$. Suppose that the expected value of $\underline{\delta}_j$ is given by

$$E\,\underline{\delta}_j = \boldsymbol{z}_{2j}\boldsymbol{\gamma}$$

for $1 \times u$ vectors $\boldsymbol{z}_{2j}$ and a regression coefficient $\boldsymbol{\gamma}$. Accordingly, $\underline{\delta}_j$ is written as $\underline{\delta}_j = \boldsymbol{z}_{2j}\boldsymbol{\gamma} + \underline{\tilde{\delta}}_j$. Note that a term in $\mathbf{1}_{n_j}\,E\,\underline{\delta}_j$ which is a linear combination of $\boldsymbol{X}_j$ will be absorbed into the model term $\boldsymbol{X}_j\boldsymbol{\beta}$, so this misspecification is non-trivial only if $\mathbf{1}_{n_j}\boldsymbol{z}_{2j}$ cannot be written as a linear combination $\boldsymbol{X}_j\boldsymbol{A}$ for some weight matrix $\boldsymbol{A}$ independent of j.

The question now is in the first place, how the parameter estimates are affected by the incorrectness of the assumption that $\underline{\delta}_j$ has a zero expected

value, corresponding to the omission of the term $\boldsymbol{z}_{2j}\boldsymbol{\gamma}$ from the model equation.

It is useful to split the variable $\boldsymbol{X}_j$ into its cluster mean $\bar{\boldsymbol{X}}_j$ and the within-cluster deviation variable $\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j - \bar{\boldsymbol{X}}_j$:

$$\boldsymbol{X}_j = \bar{\boldsymbol{X}}_j + \tilde{\boldsymbol{X}}_j$$

where

$$\bar{\boldsymbol{X}}_j = \mathbf{1}_{n_j}(\mathbf{1}'_{n_j}\mathbf{1}_{n_j})^{-1}\mathbf{1}'_{n_j}\boldsymbol{X}_j.$$

Then the data-generating model can be written as

$$\underline{\boldsymbol{y}}_j = \bar{\boldsymbol{X}}_j\boldsymbol{\beta} + \tilde{\boldsymbol{X}}_j\boldsymbol{\beta} + \mathbf{1}_{n_j}\boldsymbol{z}_{2j}\boldsymbol{\gamma} + \mathbf{1}_{n_j}\underline{\tilde{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j \ ,$$

for random effects $\underline{\tilde{\delta}}_j$ which do satisfy the condition that they have zero expected values.

A bias in the estimation of $\boldsymbol{\beta}$ will be caused by lack of orthogonality of the matrices $\boldsymbol{X}_j = \bar{\boldsymbol{X}}_j + \tilde{\boldsymbol{X}}_j$ and $\mathbf{1}_{n_j}\boldsymbol{z}_{2j}$. Since the definition of $\tilde{\boldsymbol{X}}_j$ implies that $\tilde{\boldsymbol{X}}_j$ is orthogonal to $\mathbf{1}_{n_j}\boldsymbol{z}_{2j}$, it is clear that $\bar{\boldsymbol{X}}_j$ is the villain of the piece: analogous to the situation of a misspecified General Linear Model, there will be a bias if the cluster mean of $\boldsymbol{X}$ is non-zero, $\bar{\boldsymbol{X}}'_j\mathbf{1}_{n_j} \neq 0$. If it is non-zero, there is an obvious solution: extend the fixed part by giving separate fixed parameters $\boldsymbol{\beta}_1$ to the cluster means $\bar{\boldsymbol{X}}$ and $\boldsymbol{\beta}_2$ to the deviation variables $\tilde{\boldsymbol{X}}$, so that the working model reads

$$\underline{\boldsymbol{y}}_j = \bar{\boldsymbol{X}}_j\boldsymbol{\beta}_1 + \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_2 + \mathbf{1}_{n_j}\underline{\delta}_j + \underline{\boldsymbol{\epsilon}}_j$$

(taking out the zero columns from $\bar{\boldsymbol{X}}_j$ and $\tilde{\boldsymbol{X}}_j$, which are generated by columns in $\boldsymbol{X}_j$ which themselves are within-cluster deviation variables or level-two variables, respectively). An equivalent working model is obtained by adding to (3.1) the fixed effects of the non-constant cluster means $\bar{\boldsymbol{X}}_j$. In this way, the bias in the fixed effect estimates due to ignoring the term $\boldsymbol{z}_{2j}\boldsymbol{\gamma}$ is absorbed completely by the parameter estimate for $\boldsymbol{\beta}_1$, and this misspecification does not affect the unbiasedness of the estimate for $\boldsymbol{\beta}_2$. The estimate for the level-2 variance $\mathrm{Var}(\underline{\delta}_j)$ will be affected, which is inescapable if there is no knowledge about $\boldsymbol{z}_{2j}$, but the estimate for the level-1 variance σ^2 will be consistent.

In the practice of multilevel analysis, it is known that the cluster means often have a substantively meaningful interpretation, different from the level-one variables from which they are calculated (cf. the discussion in sections 3.6 and 4.5 of Snijders and Bosker [51] about within- and between-group regressions). This often leads to a substance-matter related rationale for including the cluster means among the variables with fixed effects.

It can be concluded that in a two-level random intercept model, the sensitive part of the assumption that the level-two random effects have a zero expected value, is the orthogonality of these expected values to the cluster means of the variables $\boldsymbol{X}$ with fixed effects. This orthogonality can be tested simply by testing the effects of these cluster means included as additional variables in the fixed part of the model. This can be interpreted as testing the equality between the within-cluster regression coefficient and the between-cluster coefficient. This test — or at least a test with the same purpose — is often referred to as the Hausman test. (Hausman [26] proposed a general procedure for tests of model specification, of which the test for equality of the within-cluster and between-cluster coefficients is an important special case. Also see Baltagi [3], who shows on p. 69 that this case of the Hausman test is equivalent to testing the effect of the cluster means $\bar{\boldsymbol{X}}$.)

In econometrics, the Hausman test for the difference between the within-cluster and between-cluster regression coefficients is often seen as a test for deciding whether to use a random or fixed coefficient model for the level-two residuals $\boldsymbol{\delta}_j$. The preceding discussion shows that this is slightly beside the point. If there is a difference between the within-cluster and between-cluster regression coefficients, which is what this Hausman test intends to detect, then unbiased estimates for the fixed within-cluster effects can be obtained also with random coefficient models, provided that the cluster means of the explanatory variables are included among the fixed effect variables $\boldsymbol{X}$. Including the cluster means will lead to an increase of the number of fixed effects by at most r, which normally is much less than the $m-1$ fixed effects required for including fixed main effects of the clusters. Whether or not to use a random coefficient model depends on other considerations, as discussed earlier in this section. Fielding [16] gives an extensive discussion of this issue, and warns against the oversimplification of using this Hausman test without further thought to decide between random effects and fixed effects models.

Now consider the general case that $\boldsymbol{Z}$ has some arbitrary positive dimension s. Let the expected value of the level-two random effects $\underline{\boldsymbol{\delta}}_j$ in the data-generating model be given by

$$E\,\underline{\boldsymbol{\delta}}_j = \boldsymbol{Z}_{2j}\boldsymbol{\gamma},$$

instead of the assumed value of $\boldsymbol{\emptyset}$. It may be assumed that $\boldsymbol{Z}_j\boldsymbol{Z}_{2j}$ cannot be expressed as a linear combination $\boldsymbol{X}_j\boldsymbol{A}$ for some matrix $\boldsymbol{A}$ independent of j, because otherwise the contribution caused by $E\,\underline{\boldsymbol{\delta}}_j$ could be absorbed into $\boldsymbol{X}_j\boldsymbol{\beta}$.

Both $\boldsymbol{X}_j$ and $\boldsymbol{y}_j$ are split in two terms, the within-cluster projections $\vec{\boldsymbol{X}}_j$ and $\vec{\boldsymbol{y}}_j$ on the linear space spanned by the variables $\boldsymbol{Z}_j$,

$$\vec{\boldsymbol{X}}_j = \boldsymbol{Z}_j(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1}\boldsymbol{Z}_j'\boldsymbol{X}_j \quad \text{and} \quad \vec{\boldsymbol{y}}_j = \boldsymbol{Z}_j(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1}\boldsymbol{Z}_j'\boldsymbol{y}_j\,,$$

and the difference variables

$$\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j - \vec{\boldsymbol{X}}_j \quad \text{and} \quad \tilde{\boldsymbol{y}}_j = \boldsymbol{y}_j - \vec{\boldsymbol{y}}_j \,.$$

The projection $\vec{\boldsymbol{X}}_j$ can be regarded as the prediction of $\boldsymbol{X}_j$, produced by the ordinary least squares (OLS) regression of $\boldsymbol{X}_j$ on $\boldsymbol{Z}_j$ for cluster j separately, and the same for $\vec{\boldsymbol{y}}_j$. The data-generating model now is written as

$$\underline{\boldsymbol{y}}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma} + \boldsymbol{Z}_j\underline{\tilde{\boldsymbol{\delta}}}_j + \underline{\boldsymbol{\epsilon}}_j \ ,$$

where again the $\underline{\tilde{\boldsymbol{\delta}}}_j$ do have zero expected values.

The distribution of $\underline{\boldsymbol{y}}_j$ is the multivariate normal

$$\underline{\boldsymbol{y}}_j \sim \mathcal{N}(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}, \boldsymbol{V}_j),$$

where

$$\boldsymbol{V}_j = \sigma^2\boldsymbol{I}_{n_j} + \boldsymbol{Z}_j\boldsymbol{\Omega}(\boldsymbol{\xi})\boldsymbol{Z}_j' \,. \tag{3.4}$$

Hence the log-likelihood function of the data-generating model is given by

$$-\tfrac{1}{2}\sum_j \Big(\log\det(\boldsymbol{V}_j) + (\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma})'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma})\Big).$$

The inverse of $\boldsymbol{V}_j$ can be written as [41, 44]

$$\boldsymbol{V}_j^{-1} = \sigma^{-2}\boldsymbol{I}_{n_j} - \boldsymbol{Z}_j\boldsymbol{A}_j\boldsymbol{Z}_j' \,, \tag{3.5}$$

for a matrix

$$\boldsymbol{A}_j = \sigma^{-2}(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1} - (\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1}\big(\sigma^2(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1} + \boldsymbol{\Omega}(\boldsymbol{\xi})\big)^{-1}(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1}.$$

This implies that

$$\begin{aligned}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma})'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}) \\ = \big(\vec{\boldsymbol{y}}_j - \vec{\boldsymbol{X}}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}\big)'\boldsymbol{V}_j^{-1}\big(\vec{\boldsymbol{y}}_j - \vec{\boldsymbol{X}}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}\big) \\ + \sigma^{-2}\|\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}\|^2,\end{aligned}$$

where $\|\cdot\|$ denotes the usual Euclidean norm. The log-likelihood is

$$\begin{aligned}-\tfrac{1}{2}\sum_j \Big(&\log\det(\boldsymbol{V}_j) \\ &+ \big(\vec{\boldsymbol{y}}_j - \vec{\boldsymbol{X}}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}\big)'\boldsymbol{V}_j^{-1}\big(\vec{\boldsymbol{y}}_j - \vec{\boldsymbol{X}}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}\big) \\ &+ \sigma^{-2}\|\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}\|^2\Big).\end{aligned} \tag{3.6}$$

This shows that the omission from the model of $\boldsymbol{Z}_j\boldsymbol{Z}_{2j}\boldsymbol{\gamma}$ will affect the estimates only through the term $\vec{\boldsymbol{X}}_j\boldsymbol{\beta}$. If now separate fixed parameters are given to $\vec{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}$ so that the working model is

$$\underline{\boldsymbol{y}}_j = \vec{\boldsymbol{X}}_j\boldsymbol{\beta}_1 + \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_2 + \boldsymbol{Z}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j\,,$$

the bias due to neglecting the term $\boldsymbol{Z}_{2j}\boldsymbol{\gamma}$ in the expected value of $\underline{\boldsymbol{\delta}}_j$ will be absorbed into the estimate of $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ will be an unbiased estimate for the fixed effect of $\boldsymbol{X}$. The log-likelihood (3.6) shows that the ML and REML estimates of $\boldsymbol{\beta}_2$ are equal to the OLS estimate based on the deviation variables $\vec{\boldsymbol{y}}_j$, and also equal to the OLS estimate in the model obtained by replacing the random effects $\underline{\boldsymbol{\delta}}_j$ by fixed effects.

This discussion shows that in the general case, if one is uncertain about the validity of the condition that the level-two random effects have zero expected values, and one wishes to retain a random effects model rather than work with a model with a large number (viz., ms) of fixed effects, it is advisable to add to the model the fixed effects of the variables

$$\vec{\boldsymbol{X}}_j = \boldsymbol{Z}_j(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1}\boldsymbol{Z}_j'\boldsymbol{X}_j\,, \tag{3.7}$$

i.e., the predictions of the variables in $\boldsymbol{X}$ by within-cluster OLS regression of $\boldsymbol{X}_j$ on $\boldsymbol{Z}_j$. The model term $\boldsymbol{Z}_j\,E\,\underline{\boldsymbol{\delta}}_j$ will be entirely absorbed into the fixed effects of $\vec{\boldsymbol{X}}_j$, and the estimates of $\boldsymbol{\beta}_2$ will be unbiased for the corresponding elements of $\boldsymbol{\beta}$ in (3.1). Depending on the substantive context, there may well be a meaningful interpretation of the constructed level-two variables (3.7).

3.3 Residuals

Like in other regression-type models, residuals (which term now is used also to refer to estimates of the residuals $\underline{\boldsymbol{\epsilon}}$ and $\underline{\boldsymbol{\delta}}$ in (3.1)) play an important exploratory role for model checking in multilevel models. For each level there is a set of residuals and a residual analysis can be executed. One of the practical questions is, whether residual checking should be carried out upward — starting with level one, then continuing with level two, etc. — or downward — starting from the highest level and continuing with each subsequent lower level. The literature contains different kinds of advice. For example, Raudenbush and Bryk [45] suggest an upward approach for model construction, whereas Langford and Lewis [35] propose a downward approach for the purpose of outlier inspection. In our view, the argument given by Hilden-Minton [27] is convincing: level-one residuals can be studied unconfounded by the higher-level residuals, but the reverse is impossible. Therefore, the upward approach is preferable for the careful checking of model assumptions. However, if one wishes to carry out a quick check for outliers, a downward approach may be very efficient.

This section first treats the 'internal' standardization of the residuals. Externally standardized residuals, also called deletion residuals, are treated in section 3.3.5.

3.3.1 Level-One Residuals

In this section we assume that level-one residuals are i.i.d. Residuals at level one which are unconfounded by the higher-level residuals can be obtained, as remarked by Hilden-Minton [27], as the OLS residuals calculated separately within each level-two cluster. These are just the same as the estimated residuals in the OLS analysis of the fixed effects model, where all level-two (or higher-level, if there are any higher levels) residuals are treated as fixed rather than random. These will be called here the OLS within-cluster residuals. Consider again model (3.1) with the further specification (3.1d). When $\check{\boldsymbol{X}}_j$ is the matrix containing all non-redundant columns in $(\boldsymbol{X}_j \ \boldsymbol{Z}_j)$ and $\boldsymbol{P}_j$ is the corresponding projection matrix (the "hat matrix")

$$\boldsymbol{P}_j = \check{\boldsymbol{X}}_j \big(\check{\boldsymbol{X}}_j' \check{\boldsymbol{X}}_j\big)^{-1} \check{\boldsymbol{X}}_j' \,,$$

the OLS within-cluster residuals are given by

$$\hat{\underline{\boldsymbol{\epsilon}}}_j = \Big(\boldsymbol{I}_{n_j} - \boldsymbol{P}_j\Big)\, \underline{\boldsymbol{y}}_j \,.$$

The model definition implies that

$$\hat{\underline{\boldsymbol{\epsilon}}}_j = \Big(\boldsymbol{I}_{n_j} - \boldsymbol{P}_j\Big)\, \underline{\boldsymbol{\epsilon}}_j \,, \tag{3.8}$$

which shows that indeed these residuals depend only on the level-one residuals $\underline{\boldsymbol{\epsilon}}_j$ without confounding by the level-two residuals $\underline{\boldsymbol{\delta}}_j$.

These level-one residuals can be used for two main purposes. In the first place, for investigating the specification of the within-cluster model, i.e., the choice of the explanatory variables contained in $\boldsymbol{X}$ and $\boldsymbol{Z}$. Linearity of the dependence on these variables can be checked by plotting the residuals $\hat{\underline{\boldsymbol{\epsilon}}}_j$ against the variables in $\boldsymbol{X}$ and $\boldsymbol{Z}$. The presence of outliers and potential effects of omitted but available variables can be studied analogously.

In the second place, the homoscedasticity assumption (3.1d) can be checked. Equation (3.8) implies that, if the model assumptions are correct,

$$\hat{\underline{\epsilon}}_{ij} \sim \mathcal{N}\big(0, \sigma^2(1 - h_{ij})\big) \tag{3.9}$$

where h_{ij} is the i-th diagonal element of the hat matrix $\boldsymbol{P}_j$. This implies that the "semi-standardized residuals"

$$\check{\underline{\epsilon}}_{ij} = \frac{\hat{\underline{\epsilon}}_{ij}}{\sqrt{1 - h_{ij}}}$$

have a normal distribution with mean 0 and variance σ^2. For checking homoscedasticity, the squared semi-standardized residuals can be plotted against explanatory variables or in a meaningful order. This is informative only under

the assumption that the expected value of the residuals is indeed 0. Therefore these heteroscedasticity checks should be performed only after having ascertained the linear dependence of the fixed part on the explanatory variables.

To check linearity and homoscedasticity as a function of explanatory variables, if the plot of the residuals just shows a seemingly chaotic mass of scatter, it often is helpful to smooth the plots of residuals against explanatory variables, e.g., by moving averages or by spline smoothers. We find it particularly helpful to use smoothing splines [cf. 21], choosing the smoothing parameter so as to minimize the cross-validatory estimated prediction error.

If there is evidence of inhomogeneity of level-one variances, the level-one model is in doubt and attempts to improve it are in order. The analysis of level-one residuals might suggest non-linear transformations of the explanatory variables, as discussed in the second half of this chapter, or a heteroscedastic level-one model. Another possibility is to apply a non-linear transformation to the dependent variable. Atkinson [2] has an illuminating discussion of non-linear transformations of the dependent variable in single-level regression models. Hodges [28, p. 506] discusses Box-Cox transformations for multilevel models.

As an example, consider the data set provided with the MLwiN software [19] in the worksheet `tutorial.ws`. This includes data for 4059 students in 65 schools; we use the normalized exam score (`normexam`) (mean 0, variance 1) as the dependent variable and only the standardized reading test (`standlrt`) as an explanatory variable. The two mentioned uses of the OLS level-one residuals will be illustrated.

Table 3.1. Parameter estimates for models fitted to normalized exam scores.

	Model 1		Model 2		Model 3	
Fixed part						
constant term	0.002	(.040)	−0.017	(.041)	−0.017	(.041)
`standlrt`	0.563	(.012)	0.604	(.021)	0.605	(.021)
`standlrt`^{2}			0.017	(.009)	0.017	(.008)
`standlrt`^{3}			−0.013	(.005)	−0.013	(.005)
Random part						
Level 2: ω_{11}	0.092	(.018)	0.093	(.018)	0.095	(.019)
Level 1: σ^2	0.566	(.013)	0.564	(.013)	0.564	(.013)
Level 1: θ_2					−0.007	(.003)
deviance	9357.2		9346.2		9341.4	

When the OLS within-cluster residuals are plotted against the explanatory variable `standlrt`, an unilluminating cloud of points is produced. Therefore only the smoothed residuals are plotted in Figure 3.1.

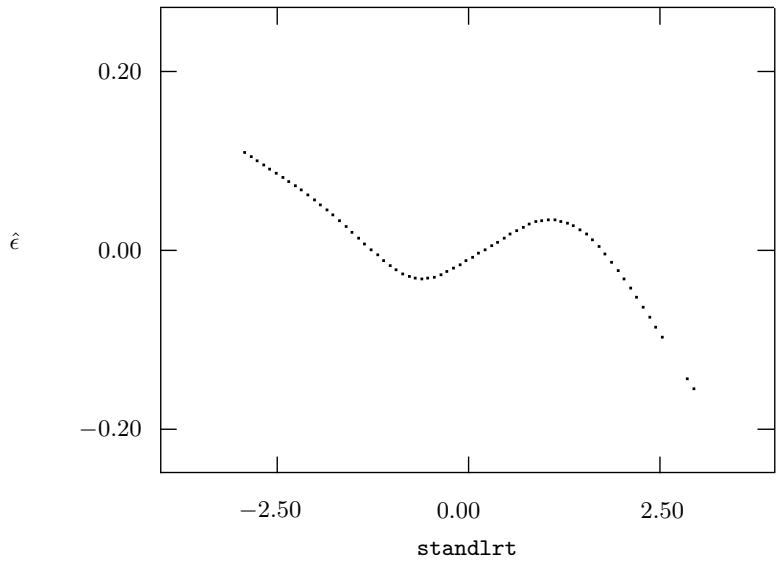

Fig. 3.1. Smoothing spline approximation for OLS within-cluster residuals ($\hat{\epsilon}$) under Model 1 against standardized reading test (`standlrt`).

This figure shows a smooth curve suggestive of a cubic polynomial. The shape of the curve suggests to include the square and cube of `standlrt` as extra explanatory variables. The resulting model estimates are presented as Model 2 in Table 3.1. Indeed the model improvement is significant ($\chi^2 = 11.0$, d.f. $= 2$, $p < .005$).

As a check of the level-one homoscedasticity, the semi-standardized residuals (3.9) are calculated for Model 2. The smoothed squared semi-standardized residuals are plotted against `standlrt` in Figure 3.2.

This figure suggests that the level-one variance decreases linearly with the explanatory variable. A model with this specification (cf. section 3.2.1),

$$\mathrm{Var}(\underline{\epsilon}_{ij}) = \sigma^2 + \theta_2 \, \mathtt{standlrt}_{ij} \ ,$$

is presented as Model 3 in Table 3.1. The heteroscedasticity is a significant model improvement ($\chi^2 = 4.8$, d.f. $= 1$, $p < .05$).

3.3.2 Homogeneity of Variance across Clusters

The OLS within-cluster residuals can also be used in a test of the assumption that the level-one variance is the same in all level-two units against the specific alternative hypothesis that the level-one variance varies across the level-two units. Formally, this means that the null hypothesis (3.1d) is tested against the alternative

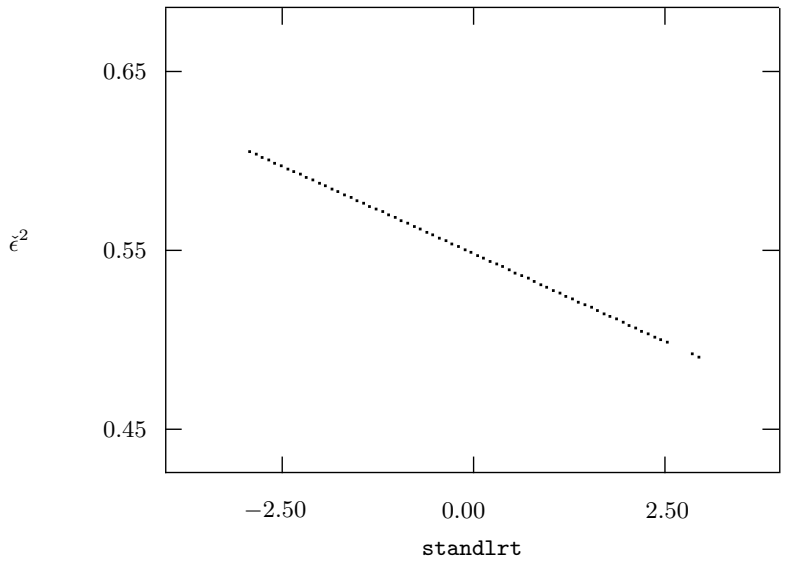

Fig. 3.2. Smoothing spline approximation for the squared semi-standardized OLS within-cluster residuals ($\check{\epsilon}^2$) under Model 2 against the standardized reading test (`standlrt`).

$$\boldsymbol{\Sigma}_j(\boldsymbol{\theta}) = \sigma_j^2 \boldsymbol{I}_{n_j} ,$$

where the σ_j^2 are unspecified and not identical.

Indicating the rank of $\check{\boldsymbol{X}}_j$ defined in section 3.3.1 by r_j, the within-cluster residual variance is

$$\underline{s}_j^2 = \frac{1}{n_j - r_j} \hat{\boldsymbol{\epsilon}}_j' \hat{\boldsymbol{\epsilon}}_j \ .$$

If model (3.1d) is correct, $(n_j - r_j)\underline{s}_j^2/\sigma^2$ has a chi-squared distribution with $(n_j - r_j)$ degrees of freedom. The homogeneity test of Bartlett and Kendall [4] can be applied here (it is also proposed in Raudenbush and Bryk [45, p. 264] and Snijders and Bosker [51, p. 127]). Denoting $\sum n_j = n_+$, $\sum r_j = r_+$ and

$$\underline{ls}_{\text{pooled}} = \frac{1}{n_+ - r_+} \sum_j (n_j - r_j) \log(\underline{s}_j^2), \tag{3.10}$$

the test statistic is given by

$$\underline{H} = \sum_j \frac{n_j - r_j}{2} \left(\log(\underline{s}_j^2) - \underline{ls}_{\text{pooled}} \right)^2 . \tag{3.11}$$

Under the null hypothesis this statistic has approximately a chi-squared distribution with $\tilde{m} - 1$ degrees of freedom, where $\tilde{m}$ is the number of clusters

included in the summation (this could be less than m because some small clusters might be skipped).

This chi-squared approximation is valid if the degrees of freedom $n_j - r_j$ are large enough. If this approximation is in doubt, a Monte Carlo test can be used. This test is based on the property that, under the null hypothesis, $(n_j - r_j)\underline{s}_j^2/\sigma^2$ has an exact chi-squared distribution, and the unknown parameter σ^2 does not affect the distribution of $\underline{H}$ because its contribution in (3.11) cancels out. This implies that under the null hypothesis the distribution of $\underline{H}$ does not depend on any unknown parameters, and a random sample from its distribution can be generated by randomly drawing random variables $\underline{c}_j^2$ from chi-squared distributions with $(n_j - r_j)$ d.f. and applying formulae (3.10) and (3.11) to $\underline{s}_j^2 = \underline{c}_j^2/(n_j - r_j)$. By simulating a sufficiently large sample from the null distribution of $\underline{H}$, the p-value of an observed value can be approximated to any desired precision.

3.3.3 Level-Two Residuals

There are two main ways for predicting[4] the level-two residuals $\underline{\boldsymbol{\delta}}_j$: the OLS method (based on treating them as fixed effects $\boldsymbol{\delta}_j$) and the empirical Bayes (EB) method. The empirical Bayes 'estimate' of $\underline{\boldsymbol{\delta}}_j$ can be defined as its conditional expected value given the observations $\underline{\boldsymbol{y}}_1, \ldots, \underline{\boldsymbol{y}}_m$, plugging in the parameter estimates for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\xi}$. (In the name, 'Bayes' refers to the conditional expectation and 'empirical' to plugging in the estimates.)

The advantage of the EB method is that it is more precise, but the disadvantage is its stronger dependence on the model assumptions. The two approaches were compared by Waternaux et al. [59] and Hilden-Minton [27]. Their conclusion was that, provided the level-one model (i.e., the assumptions about the level-one predictors included in $\boldsymbol{X}$ and about the level-one residuals $\underline{\boldsymbol{\epsilon}}_j$) is adequate, it is advisable to use the EB estimates.

Basic properties of the multivariate normal distribution imply that the EB level-two residuals are given by

$$\begin{aligned}
\hat{\underline{\boldsymbol{\delta}}}_j &= E\{\underline{\boldsymbol{\delta}}_j \mid \underline{\boldsymbol{y}}_1, \ldots, \underline{\boldsymbol{y}}_m\} \qquad \text{(using parameter estimates } \hat{\underline{\boldsymbol{\beta}}}, \hat{\boldsymbol{\theta}}, \hat{\underline{\boldsymbol{\xi}}}) \\
&= \hat{\underline{\boldsymbol{\Omega}}}\boldsymbol{Z}_j'\hat{\underline{\boldsymbol{V}}}_j^{-1}(\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j\hat{\underline{\boldsymbol{\beta}}}) \\
&= \hat{\underline{\boldsymbol{\Omega}}}\boldsymbol{Z}_j'\hat{\underline{\boldsymbol{V}}}_j^{-1}\big(\boldsymbol{Z}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j - \boldsymbol{X}_j(\hat{\underline{\boldsymbol{\beta}}} - \boldsymbol{\beta})\big)
\end{aligned}$$

where

[4] Traditional statistical terminology is to reserve the word 'estimation' for empirical ways to obtain reasonable values for parameters, and use 'prediction' for ways to empirically approximate unobserved outcomes of random variables. We shall not consistently respect this terminology, since almost everybody writes about *estimation* of residuals.

$$\boldsymbol{V}_j = \mathrm{Cov}(\underline{\boldsymbol{y}}_j) = \boldsymbol{Z}_j \boldsymbol{\Omega} \boldsymbol{Z}_j' + \boldsymbol{\Sigma}_j \,, \tag{3.12a}$$

$$\underline{\hat{\boldsymbol{V}}}_j = \boldsymbol{Z}_j \underline{\hat{\boldsymbol{\Omega}}} \boldsymbol{Z}_j' + \underline{\hat{\boldsymbol{\Sigma}}}_j \,, \tag{3.12b}$$

with $\underline{\hat{\boldsymbol{\Omega}}} = \boldsymbol{\Omega}(\underline{\hat{\boldsymbol{\xi}}})$ and $\underline{\hat{\boldsymbol{\Sigma}}}_j = \boldsymbol{\Sigma}_j(\underline{\hat{\boldsymbol{\theta}}})$.

Some more insight into the properties of these estimated residuals may be obtained by defining the estimated reliability matrix

$$\underline{\hat{\boldsymbol{R}}}_j = \underline{\hat{\boldsymbol{\Omega}}} \boldsymbol{Z}_j' \underline{\hat{\boldsymbol{V}}}_j^{-1} \boldsymbol{Z}_j \,.$$

This matrix is the multivariate generalization of the reliability of estimation of $\underline{\boldsymbol{\delta}}_{jq}$, the ratio of the true variance of $\underline{\boldsymbol{\delta}}_{jq}$ to the variance of its OLS estimator based on cluster j (not taking into account the component of variability due to the estimation of $\boldsymbol{\beta}$), as defined by Raudenbush and Bryk [45, p. 49].

The EB residuals can be expressed as

$$\underline{\hat{\boldsymbol{\delta}}}_j = \underline{\hat{\boldsymbol{R}}}_j \underline{\boldsymbol{\delta}}_j + \underline{\hat{\boldsymbol{\Omega}}} \boldsymbol{Z}_j' \underline{\hat{\boldsymbol{V}}}_j^{-1} \underline{\boldsymbol{\epsilon}}_j - \underline{\hat{\boldsymbol{\Omega}}} \boldsymbol{Z}_j' \underline{\hat{\boldsymbol{V}}}_j^{-1} \boldsymbol{X}_j (\underline{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}). \tag{3.13}$$

The first term can be regarded as a shrinkage transform of $\underline{\boldsymbol{\delta}}_j$, the second term is the confounding due to the level-one residuals $\underline{\boldsymbol{\epsilon}}_j$, and the third term is the contribution due to the estimation of the fixed parameters $\boldsymbol{\beta}$.

Ignoring the contribution to the variances and covariances due to the estimation of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, the covariance matrix of the EB residuals is

$$\mathrm{Cov}\big(\underline{\hat{\boldsymbol{\delta}}}_j\big) = \boldsymbol{\Omega} \boldsymbol{Z}_j' \boldsymbol{V}_j^{-1} \left(\boldsymbol{V}_j - \boldsymbol{X}_j \Big(\sum_{\ell=1}^{m} \boldsymbol{X}_\ell' \boldsymbol{V}_\ell^{-1} \boldsymbol{X}_\ell \Big)^{-1} \boldsymbol{X}_j' \right) \boldsymbol{V}_j^{-1} \boldsymbol{Z}_j \boldsymbol{\Omega}. \tag{3.14}$$

The second term in the large parentheses is due to the third term in (3.13) and will be negligible if the number m of clusters is large. The resulting simpler expression is

$$\mathrm{Cov}\big(\underline{\hat{\boldsymbol{\delta}}}_j\big) \approx \boldsymbol{\Omega} \boldsymbol{Z}_j' \boldsymbol{V}_j^{-1} \boldsymbol{Z}_j \boldsymbol{\Omega}. \tag{3.15}$$

Another relevant covariance matrix contains the variances and covariances of the prediction errors. The same approximation leading to (3.15) yields

$$\mathrm{Cov}\big(\underline{\hat{\boldsymbol{\delta}}}_j - \underline{\boldsymbol{\delta}}_j\big) \approx \boldsymbol{\Omega} - \boldsymbol{\Omega} \boldsymbol{Z}_j' \boldsymbol{V}_j^{-1} \boldsymbol{Z}_j \boldsymbol{\Omega}. \tag{3.16}$$

If all n_j become very large, (3.16) tends to $\emptyset$. Expression (3.15) is the asymptotic covariance matrix for fixed n_j, which tends to $\boldsymbol{\Omega}$ if n_j tends to infinity. The variances in (3.14) and (3.15) are relevant for diagnosing properties of the residuals $\underline{\boldsymbol{\delta}}_j$ and are called *diagnostic variances* by Goldstein [18]. The variances in (3.16) are relevant for comparing residuals $\underline{\boldsymbol{\delta}}_j$ and are called *comparative* (or *conditional*) *variances*.

It may be noted that the predictions $\underline{\hat{\boldsymbol{\delta}}}_j$ are necessarily uncorrelated with the errors $(\underline{\hat{\boldsymbol{\delta}}}_j - \underline{\boldsymbol{\delta}}_j)$, because otherwise a better prediction could be made. This implies

$$\mathrm{Cov}\big(\underline{\boldsymbol{\delta}}_j\big) = \mathrm{Cov}\big(\underline{\boldsymbol{\delta}}_j - \underline{\hat{\boldsymbol{\delta}}}_j\big) + \mathrm{Cov}\big(\underline{\hat{\boldsymbol{\delta}}}_j\big),$$

which indeed is evident from the formulae.

For each of the s level-two random effects separately, various diagnostic plots can be made. The explanation of the level-two random effects by level-two variables, as reflected by the fixed main effects of level-two variables and their cross-level interaction effects with the variables contained in $\boldsymbol{Z}$, can be diagnosed for linearity by plots of the raw residuals $\underline{\hat{\boldsymbol{\delta}}}_j$ against the level-two explanatory variables. The normality and homoscedasticity assumptions for $\underline{\boldsymbol{\delta}}_j$ can be checked by normal probability plots for the s residuals separately, standardized by dividing them by the diagnostic standard deviations obtained as the square roots of the diagonal elements of (3.14) or (3.15), and by plotting the squares of these standardized residuals against the level-two variables. Such plots were proposed and discussed by Lange and Ryan [34]. Examples of these plots are given in Goldstein [18], Snijders and Bosker [51], and Lewis and Langford [37].

Eberly and Thackeray [13] showed that it is very well possible that, when such a plot shows deviations from normality, the cause is a misspecification of the fixed effects model rather than of the distribution of the random effects. This is in accordance with the general caveat that different aspects of the specification of statistical models are entwined, and the particular importance of this issue for assessing fit of multilevel models. It also supports the principle to first try achieve a good specification of the level-one model, and assess the level-two specification only after this has been done.

A diagnostic for the entire vector of level-two residuals for cluster j can be based on the standardized value

$$\underline{\hat{\boldsymbol{\delta}}}_j' \Big\{ \widehat{\mathrm{Cov}}(\underline{\hat{\boldsymbol{\delta}}}_j) \Big\}^{-1} \underline{\hat{\boldsymbol{\delta}}}_j \,. \tag{3.17}$$

If one neglects the fact that the estimated rather than the true covariance matrix is used, this statistic has a chi-squared distribution with s degrees of freedom.

With some calculations, using formula (3.5) and the approximate covariance matrix (3.15), the standardized value (3.17) is seen to be given by

$$\underline{\hat{\boldsymbol{\delta}}}_j' \Big\{ \widehat{\mathrm{Cov}}(\underline{\hat{\boldsymbol{\delta}}}_j) \Big\}^{-1} \underline{\hat{\boldsymbol{\delta}}}_j \approx \underline{\hat{\boldsymbol{\delta}}}_j^{(\mathrm{OLS})\prime} \Big(\underline{\hat{\sigma}}^2 (\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1} + \underline{\hat{\boldsymbol{\Omega}}} \Big)^{-1} \underline{\hat{\boldsymbol{\delta}}}_j^{(\mathrm{OLS})} \tag{3.18}$$

where

$$\underline{\hat{\boldsymbol{\delta}}}_j^{(\mathrm{OLS})} = (\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1} \boldsymbol{Z}_j' (\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \underline{\hat{\boldsymbol{\beta}}}_j)$$

is the OLS estimate of $\boldsymbol{\delta}_j$, estimated from the OLS within-cluster residuals $\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \underline{\hat{\boldsymbol{\beta}}}_j$. This illustrates that the standardized value can be based on the OLS residuals as well as the EB residuals, if one uses for standardization the covariance matrix $\hat{\sigma}^2(\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1} + \hat{\boldsymbol{\Omega}}$ of which the first part is the sampling

variance (level-one variance) and the second part the true variance (level-two variance) of the OLS residuals. The name of *standardized level-two residual* therefore is more appropriate for (3.18) than the name of standardized EB or OLS residual, since the latter terminology suggests a non-existing distinction.

The ordered standardized level-two residuals can be plotted against the corresponding quantiles of the chi-squared distribution with s d.f., as a check for outliers and for the multivariate normality of the level-two random effects.

3.3.4 Multivariate Residuals

The fit of the model for level-two cluster j is expressed by the multivariate residual

$$\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \underline{\hat{\boldsymbol{\beta}}}\,. \tag{3.19}$$

The covariance matrix of this residual, if we neglect the use of the estimated parameter $\underline{\hat{\boldsymbol{\beta}}}$ instead of the unknown true $\boldsymbol{\beta}$, is given by $\boldsymbol{V}_j$ in (3.12a). Accordingly, the *standardized multivariate residual* is defined by

$$\underline{M}_j^2 = (\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \underline{\hat{\boldsymbol{\beta}}})' \underline{\hat{\boldsymbol{V}}}_j^{-1} (\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \underline{\hat{\boldsymbol{\beta}}}).$$

This residual has, when the model is correct, approximately a chi-squared distribution with n_j degrees of freedom.

If all variables with fixed effects also have random effects, then $\boldsymbol{X}_j = \boldsymbol{Z}_j = \check{\boldsymbol{X}}_j$ as defined in section 3.3.1, and $r_j = r = s$. Using (3.5), it can be proved that in this case

$$\underline{M}_j^2 = (n_j - r)\,\frac{\underline{s}_j^2}{\underline{\hat{\sigma}}^2} + \underline{\hat{\boldsymbol{\delta}}}_j' \left\{ \underline{\widehat{\mathrm{Cov}}}(\underline{\hat{\boldsymbol{\delta}}}_j) \right\}^{-1} \underline{\hat{\boldsymbol{\delta}}}_j\,. \tag{3.20}$$

In words, the standardized multivariate residual (with n_j d.f.) is the sum of the scaled within-cluster residual sum of squares (with $n_j - r$ d.f.) and the standardized level-two residual (with $r = s$ d.f.). If some of the variables with fixed effects do not have a random effect, then the difference between the left-hand side and the right-hand side of (3.20) is a test statistic for the null hypothesis that the variables in $\boldsymbol{X}_j$ indeed have the effect expressed by the overall parameter estimate $\underline{\hat{\boldsymbol{\beta}}}$, i.e., the hypothesis that the variables in $\boldsymbol{X}$ and not in $\boldsymbol{Z}$ have only fixed (and not random) effects. This then approximately is a chi-squared variate with $r_j - s$ d.f.

This split implies that if the standardized multivariate residual for some cluster j is unexpectedly large, it will be informative to consider its two (or three) components and investigate whether the high value can be traced to one of these components separately.

3.3.5 Deletion Residuals

To assess the fit of the model and the possibility of outliers, it is better to calculate and standardize residuals for cluster j using parameter estimates of $\boldsymbol{\beta}$ and $\boldsymbol{V}_j$ calculated on the basis of the data set from which cluster j has been omitted. Such measures are called externally studentized residuals [11] or deletion residuals [2]. This means using the fixed parameter estimate $\hat{\underline{\boldsymbol{\beta}}}_{(-j)}$ obtained by estimating $\boldsymbol{\beta}$ from the data set from which cluster j has been omitted and estimating (3.12a) by

$$\hat{\underline{\boldsymbol{V}}}_{(-j)} = \boldsymbol{Z}_j \hat{\underline{\boldsymbol{\Omega}}}_{(-j)} \boldsymbol{Z}_j' + \hat{\underline{\boldsymbol{\Sigma}}}_{(-j)}, \tag{3.21}$$

where $\hat{\underline{\boldsymbol{\Omega}}}_{(-j)} = \boldsymbol{\Omega}(\hat{\underline{\boldsymbol{\xi}}}_{(-j)})$ and $\hat{\underline{\boldsymbol{\Sigma}}}_{(-j)} = \boldsymbol{\Sigma}_j(\hat{\underline{\boldsymbol{\theta}}}_{(-j)})$, while $\hat{\underline{\boldsymbol{\xi}}}_{(-j)}$ and $\hat{\underline{\boldsymbol{\theta}}}_{(-j)}$ are the estimates of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ based on the data set from which cluster j has been omitted.

Using these ingredients, the *deletion standardized multivariate residual* is defined by

$$\underline{M}^2_{(-j)} = \left(\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \hat{\underline{\boldsymbol{\beta}}}_{(-j)}\right)' \hat{\underline{\boldsymbol{V}}}^{-1}_{(-j)} \left(\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \hat{\underline{\boldsymbol{\beta}}}_{(-j)}\right). \tag{3.22}$$

The *deletion standardized level-two residual* (for a model where $\boldsymbol{\Sigma}_j(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{I}_{n_j}$) is defined by

$$\hat{\underline{\boldsymbol{\delta}}}^{(\mathrm{OLS})\prime}_{(-j)} \left(\hat{\underline{\sigma}}^2_{(-j)} (\boldsymbol{Z}_j' \boldsymbol{Z}_j)^{-1} + \hat{\underline{\boldsymbol{\Omega}}}_{(-j)}\right)^{-1} \hat{\underline{\boldsymbol{\delta}}}^{(\mathrm{OLS})}_{(-j)} \tag{3.23}$$

where

$$\hat{\underline{\boldsymbol{\delta}}}^{(\mathrm{OLS})}_{(-j)} = (\boldsymbol{Z}_j' \boldsymbol{Z}_j)^{-1} \boldsymbol{Z}_j' \left(\underline{\boldsymbol{y}}_j - \boldsymbol{X}_j \hat{\underline{\boldsymbol{\beta}}}_{(-j)}\right)$$

and $\hat{\underline{\sigma}}^2_{(-j)}$ is the estimate for σ^2 calculated from the data set from which cluster j was omitted.

The general idea of model diagnostics is that they should be easy, or at least quick, to compute. Elegant computational formulae have been derived for deletion residuals in the General Linear Model (see Atkinson [2]), and recently by Zewotir and Galpin [63] and Haslett and Dillane [23] also for random coefficient models with uncorrelated random coefficients. This yields the possibility of quick calculations of level-two deletion residuals. In the HLM the assumption of uncorrelated higher-level residuals is trivially satisfied for the random intercept model where $\underline{\boldsymbol{\delta}}_j$ is a column vector, but not if there are random slopes. Therefore these formulae are not generally applicable for random slope models.

Re-estimation of a multilevel model for a lot of different data sets, as implied by the definition of deletion residuals, is not very attractive from the point of view of quick computations. Two alternatives to full computation have been proposed in the literature: Lesaffre and Verbeke [36] proposed

influence statistics using an analytic approximation based on second-order Taylor expansions, and Snijders and Bosker [51] proposed a computational approximation based on a one-step estimator. The latter approximation will be followed here because of its simple generalizability to other situations. This approximation is defined as follows.

An iterative estimation algorithm is used, viz., Fisher scoring or (R)IGLS. The initial value for the estimation algorithm is the estimate obtained from the full data set. The one-step estimate is the result of a single step of the algorithm, using the data set reduced by omitting all data for cluster j. It is known from general statistical theory that such one-step estimates are asymptotically efficient. They can be quickly estimated by software that implements Fisher scoring or (R)IGLS. Therefore, all estimates denoted here with the suffix (-j) can be implemented as such one-step estimates obtained with the full-data estimate as the initial value.

3.4 Influence Diagnostics of Higher-Level Units

Next to the direct study of residuals as proposed in the previous section, another approach to model checking is to investigate the influence of individual data points, or sets of data points, on the parameter estimates. In OLS regression, the most widely known technique in this approach is Cook's distance, explained, e.g., in Cook and Weisberg [11], Atkinson [2], and Weisberg [60]. A natural way of performing such checks in multilevel models is to investigate the separate influence of each higher-level unit. This means that the estimates obtained from the total data set are compared to the estimates obtained from the data set from which a particular higher-level unit is omitted.

An influence measure of level-two unit j on the estimation of the parameters should reflect the importance of the influence of the data for this unit on the parameter estimates. First consider the regression coefficients $\boldsymbol{\beta}$. Recall that $\underline{\hat{\boldsymbol{\beta}}}$ is the estimate obtained from the full data set, and $\underline{\hat{\boldsymbol{\beta}}}_{(-j)}$ the estimate obtained from the data set from which unit j has been omitted, or an approximation to this estimate. The difference between these two estimates should be standardized on the basis of the inherent imprecision expressed by the covariance matrix of these estimates. In Lesaffre and Verbeke [36] and Snijders and Bosker [51] it was proposed to use the estimated covariance matrix of the estimators obtained from the full data set. Since the diagnostic measure has the aim to detect unduly influential units, it should be taken into account, however, that the unit under scrutiny also might have an undue influence on this estimated covariance matrix. Therefore it is more appropriate to use the estimated covariance matrix of the estimator obtained from the reduced data set. It may be noted that the computation of this matrix is

straightforward in the computational approach of Snijders and Bosker [51], but does not fit well in the analytic approach of Lesaffre and Verbeke [36].

Denote by $\hat{\underline{\boldsymbol{S}}}_{F(\text{-}j)}$ the estimated covariance matrix of $\hat{\underline{\boldsymbol{\beta}}}_{(\text{-}j)}$ as calculated from the data set from which level-two unit j has been omitted. Then a standardized measure of the influence of this unit on the fixed parameter estimates is

$$\underline{C}_j^F = \frac{1}{r}\left(\hat{\underline{\boldsymbol{\beta}}} - \hat{\underline{\boldsymbol{\beta}}}_{(\text{-}j)}\right)' \hat{\underline{\boldsymbol{S}}}_{F(\text{-}j)}^{-1} \left(\hat{\underline{\boldsymbol{\beta}}} - \hat{\underline{\boldsymbol{\beta}}}_{(\text{-}j)}\right). \tag{3.24}$$

This formula is analogous to Cook's distance for the General Linear Model.

For the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ of the random part of the model, the same procedure can be followed. Indicating these parameters jointly by $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\xi})$, this leads to the influence measure

$$\underline{C}_j^R = \frac{1}{p}\left(\hat{\underline{\boldsymbol{\eta}}} - \hat{\underline{\boldsymbol{\eta}}}_{(\text{-}j)}\right)' \hat{\underline{\boldsymbol{S}}}_{R(\text{-}j)}^{-1} \left(\hat{\underline{\boldsymbol{\eta}}} - \hat{\underline{\boldsymbol{\eta}}}_{(\text{-}j)}\right), \tag{3.25}$$

where the analogous definitions are used for $\hat{\underline{\boldsymbol{\eta}}}_{(\text{-}j)}$ and $\hat{\underline{\boldsymbol{S}}}_{R(\text{-}j)}$, and p is the total number of parameters in $\boldsymbol{\eta}$. Since the parameters of the fixed and random parts are asymptotically uncorrelated [40], these two influence measures can be combined in the overall influence measure

$$\underline{C}_j = \frac{1}{r+p}\left(r\underline{C}_j^F + p\underline{C}_j^R\right). \tag{3.26}$$

Comparisons with alternative definitions for diagnostics of the type of Cook's distance are given in Verbeke and Molenberghs [55] and Skrondal and Rabe-Hesketh [49].

The influence of a part of the data set on the parameter estimates depends on the *fit* of the model to this part of the data together with the *leverage* of this part, i.e., its potential to influence the parameters as determined from the amount of data and the distribution of the explanatory variables $\boldsymbol{X}$ and $\boldsymbol{Z}$. For a level-two unit, its size n_j and the distribution of $\boldsymbol{X}_j$ and $\boldsymbol{Z}_j$ determine the leverage. The fit can be measured by the deletion standardized multivariate residual (3.22). A poorly fitting cluster with small leverage will not do much damage to the results of the data analysis. If the model fits well, while there are no systematic differences between the clusters in the distribution of $\boldsymbol{X}_j$ and $\boldsymbol{Z}_j$, and the n_j are small compared to $\sum_j n_j$, the diagnostics (3.24) – (3.26) will have expected values which are roughly proportional to the cluster sizes n_j. A plot of these diagnostics against n_j may draw the attention toward clusters that have an undue influence on the parameter estimates. This information can be combined with the p-values for the deletion standardized multivariate residuals (3.22) obtained from the chi-squared distribution with n_j degrees of freedom, which give information on the fit of the clusters independently of their leverage.

3.5 Simulation-Based Assessment of Model Specification

It was shown above that the specification of the level-one model can be investigated by considering within-cluster relations between variables or, equivalently, by fixed effect models. These are analyses that effectively reduce the HLM to the General Linear Model, for which distributional properties of many statistics have been derived. These properties can be found in the ample literature of model diagnostics in such models. Properties of higher-level diagnostics cannot be derived by going back to the General Linear Model, and tend to be approximate or unknown. Longford [42] elaborates how simulations can be used to assess p-values of arbitrary statistics based e.g. on residuals or influence measures. This is done by repeatedly simulating the data under the tested model assumptions and considering the resulting distribution of the statistic under consideration; such a procedure is also called the parametric bootstrap, cf. Van der Leeden et al. [53].

Among such simulation-based procedures, the Monte Carlo test proposed at the end of section 3.3.2 illustrates the relative simplicity of checking the level-one model by the fact that the distribution of the statistic considered is independent of any unknown parameters (it is said to be *pivotal*), contrasting to the general case for higher-level diagnostics.

3.6 Non-linear Transformations in the Fixed Part

One of the purposes for which one can use the residuals discussed in the preceding sections, is to give guidance of an informal kind when investigating possible non-linear effects of explanatory variables. The remainder of this chapter presents methods to examine non-linear fixed effects of explanatory variables by incorporating them formally into the model.

We consider multilevel models for analyzing the effect of a predictor x on a response variable $\underline{y}$ under the assumption that this effect is a non-linear function $f(x)$ with an unknown functional form. The latter situation is common, e.g., when x refers to time in longitudinal studies, since the effect of time on the response is usually complex and not well understood. Then it seems sensible to approximate $f(x)$ by a flexible function that requires only minimal prior knowledge about $f(x)$ and still provides insight into the dependence between $\underline{y}$ and x.

In the following sections, we will consecutively discuss multilevel models in which the non-linear function $f(x)$ is approximated by a polynomial function, a regression spline, and a smoothing spline. As a guiding model in the discussion, we will use a two-level model for normal responses. Since longitudinal data offer the main (but not only) applications of this approach, clusters will be regarded as individual subjects, and level-one units as repeated

measurements of the subjects. We assume that the responses of subject j are generated by

$$\underline{\boldsymbol{y}}_j = f(\boldsymbol{x}_j) + \boldsymbol{X}_{2j}\boldsymbol{\beta} + \boldsymbol{Z}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j \,. \tag{3.27}$$

The difference with respect to model (3.1) is that the fixed part is split into, first, a real-valued variable $\boldsymbol{x}$ with a non-linear effect, and second, variables $\boldsymbol{X}_2$ with linear effects.

3.7 Polynomial Model

The polynomial model for multilevel data was put forward by many authors including Goldstein [17], Bryk and Raudenbush [8], and Snijders [50]. The use of a polynomial approximation seems quite natural since it can be regarded as a Taylor expansion of the true unknown function. The Q-th degree polynomial equals

$$f_{\mathrm{pol}}(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_Q x^Q \;.$$

The smoothness of $f_{\mathrm{pol}}(x)$ is controlled by the degree Q. The function $f_{\mathrm{pol}}(x)$ is a linear combination of polynomial terms x, x^2, ..., x^Q and therefore this model remains within the confines of the Hierarchical Linear Model, and can be estimated straighforwardly like any other such model. The number of parameters only depends on the degree Q so that the polynomial model is easy to estimate also when $\boldsymbol{x}_j$ differs among subjects. However, estimation problems may arise when $\boldsymbol{x}_j$ is badly scaled. In that case, a simple solution that works for many data sets is to subtract the subject mean from $\boldsymbol{x}_j$. A slightly more elaborate solution is to orthogonalize the polynomial terms using the Gram-Schmidt method.

An attractive feature of the polynomial model is that the regression coefficients can be interpreted as growth parameters which often are of substantive interest. The effect α_1, for instance, can be interpreted as the rate of change in the response at $x = 0$ which may be a useful parameter of a growth process.

The function $f(x)$ is not always well approximated by a low-degree polynomial, however. In human growth studies, for example, polynomials may fail to produce a smooth and accurate fit because of strong growth during the first year of age and early adolescence [5]. The underlying problem is that a polynomial exhibits non-local behavior which means that a change in one of the regression coefficients α_q leads to a change in the estimated $f_{\mathrm{pol}}(x)$ for (nearly) all values of x. A consequence of non-local behaviour is that when the fit at a certain value of x is improved by increasing Q, the fit may become poorer at other values of x. In general, a polynomial with a high value of Q tends to fit accurately in intervals of x with many observations but this may be achieved at the cost of a poor fit at other values of x.

3.8 Regression Spline Model

A regression spline [61] consists of piecewise polynomials that are joined at locations on the x-axis named knots. At each knot, two Q-th degree polynomials are connected such that the $(Q-1)$-th derivative of the resulting function exists and is itself a continuous function of x. A popular regression spline in practical data analysis is the cubic or third-degree regression spline, the second derivative of which is continuous at the knots. Regression splines are more flexible than polynomials and often provide a better fit in the presence of strong local non-linearity. However, regression splines are more difficult to specify than polynomials because the number of knots and the positions of the knots need to be determined. For selection of the number of knots, an ad hoc approach can be adopted in which the number of knots is increased until an accurate fit is obtained. This approach may lead to overfitting because there is no penalty for model complexity. To limit the number of knots, a possible approach is to optimize a model summary such as Akaike's Information Criterion (AIC) or the cross-validated log-likelihood [47]. Regarding the positions of the knots on the x-axis, common choices are equally spaced points or quantile points of the empirical distribution of x.

A Q-th degree regression spline with L knots at $a_1, \ldots, a_L$ can be constructed by extending a Q-th degree polynomial with L truncated polynomial terms $(x-a_l)_+^Q$ $(l = 1, \ldots, L)$, where the truncated term $(x-a_l)_+^Q$ is equal to $(x-a_l)^Q$ if $x > a_l$ and zero otherwise. The resulting function $f_{\mathrm{reg}}(x)$ can be written as

$$f_{\mathrm{reg}}(x) = \sum_{q=0}^{Q} \alpha_q x^q + \sum_{l=1}^{L} \alpha_{Q+l}(x-a_l)_+^Q \,. \tag{3.28}$$

This representation is easy to understand and the α_q's have a clear interpretation. It shows that the regression spline is a linear function of polynomial terms and therefore easy to handle, as it remains within a finite-dimensional linear function space. For numerical reasons, however, the use of truncated polynomials is not recommendable especially not when the knots are chosen close together. It often is better to work with a different set of basis functions. If $f_{\mathrm{reg}}(x)$ is a cubic regression spline, it is recommendable to write $f_{\mathrm{reg}}(x)$ as a linear combination of so-called B-splines, which are a specific set of piecewise cubic splines. Computation is stable because B-splines take nonzero values over an interval with at most five knots [12]. If $f_{\mathrm{reg}}(x)$ contains one knot at position a only (i.e., $L = 1$ in (3.28)), a simple method to improve scaling of the design matrix is to replace the term x^q in the truncated polynomial formulation of $f_{\mathrm{reg}}(x)$ by the term $(x-a)_-^q$ which equals $(x-a)^q$ if $x < a$ and 0 otherwise [51, p. 189]. Because the data columns of values of $(x-a)_-^q$ and $(x-a)_+^q$ are orthogonal, estimation is stable.

The regression spline is more flexible than the polynomial and tends to exhibit less non-local behavior. The knots are determined outside the model and good placement on the x-axis may require some trial and error. Furthermore, if only a small number of knots is used, the regression spline will not be free from non-local behavior while using too many knots is undesirable since it induces non-smooth behavior. To prevent the spline from being either non-smooth or insufficiently flexible, a possible strategy is to include a large number of knots and at the same time penalize the regression coefficients so that a smooth fit is obtained [14]. A limiting case is a function in which a knot is placed at each distinct value of x in the data set. Splines of the latter type are discussed in the next section.

3.9 Smoothing Spline Model

Suppose that the data set contains T ordered distinct values $x_1, \ldots, x_T$. The cubic smoothing spline, denoted by $f_{\mathrm{css}}(x)$, then is a cubic regression spline with knots at $x_1, \ldots, x_T$ and it is a linear function outside the interval $[x_1, x_T]$. The degree of smoothness is regulated by extending the log-likelihood function with a roughness penalty that penalizes functions for having strong curvature, that is, a large absolute second derivative $|f''_{\mathrm{css}}(x)|$. The definition of the roughness penalty is

$$-\tfrac{1}{2}\lambda \int_{x_0}^{x_{T+1}} \left\{ f''(x) \right\}^2 \, \mathrm{d}x, \tag{3.29}$$

where λ is a nonnegative smoothing parameter determining the degree of smoothing, and $x_0 < x_1$ and $x_{T+1} > x_T$.

The following basic properties of smoothing splines can be found in the literature on this topic, such as Green and Silverman [21]. The fitted cubic smoothing spline is obtained by maximizing the penalized log-likelihood, that is, the sum of the log-likelihood and the roughness penalty. An additional constraint to ensure that $f_{\mathrm{css}}(x)$ is a cubic smoothing spline does not have to be included because among all functions $f_{\mathrm{css}}(x)$ with continuous second derivatives, the unique minimizer of the penalized log-likelihood is the cubic smoothing spline. If we substitute the cubic smoothing spline $f_{\mathrm{css}}(x)$ with knots at $x_1, \ldots, x_T$ in (3.29), we can evaluate the roughness penalty as

$$-\tfrac{1}{2}\lambda \boldsymbol{f}'_{\mathrm{css}} \boldsymbol{K} \boldsymbol{f}_{\mathrm{css}} \,,$$

where $\boldsymbol{f}_{\mathrm{css}}$ is the vector of values of $f_{\mathrm{css}}(x)$ at $x_1, \ldots, x_T$. The $T \times T$ matrix $\boldsymbol{K}$ equals

$$\boldsymbol{K} = \boldsymbol{Q} \boldsymbol{R}^{-1} \boldsymbol{Q}',$$

where $\boldsymbol{Q}$ is a $T \times (T-2)$ matrix having entries $q_{i,i} = 1/(x_{i+1} - x_i)$, $q_{i+2,i} = 1/(x_{i+2} - x_{i+1})$, $q_{i+1,i} = -(q_{i,i} + q_{i+2,i})$ for $i = 1, \ldots, T-2$, and zero otherwise.

The $(T-2)\times(T-2)$ matrix $\boldsymbol{R}$ is symmetric tridiagonal with diagonal entries $r_{i,i} = \frac{1}{3}(x_{i+2} - x_i)$ for $i = 1, \ldots, T-2$. The non-zero off-diagonal entries are $r_{i,i+1} = r_{i+1,i} = \frac{1}{6}(x_{i+2} - x_{i+1})$ for $i = 1, \ldots, T-3$.

3.9.1 Estimation

The model for the responses of subject j is obtained by substituting $\boldsymbol{N}_j \boldsymbol{f}_{\text{css}}$ for $f(\boldsymbol{x}_j)$ in (3.27), where $\boldsymbol{N}_j$ is an $n_j \times T$ matrix of zeros and ones. Each row of $\boldsymbol{N}_j$ contains a single one at the entry t for which $x_{ij} = x_t$. The resulting equation is

$$\underline{\boldsymbol{y}}_j = \boldsymbol{N}_j \boldsymbol{f}_{\text{css}} + \boldsymbol{X}_{2j}\boldsymbol{\beta} + \boldsymbol{Z}_j \underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j\,, \quad j = 1, \ldots, m. \tag{3.30}$$

The model parameters to be estimated are the vector of spline values $\boldsymbol{f}_{\text{css}}$, the fixed regression coefficients $\boldsymbol{\beta}$, the level-one variance σ^2, and the level-two variance parameters $\boldsymbol{\xi}$. Given σ^2 and $\boldsymbol{\xi}$, the penalized log-likelihood is maximized by

$$\hat{\underline{\boldsymbol{f}}}_{\text{css}} = \left(\sum_{j=1}^{m} \boldsymbol{N}_j' \boldsymbol{U}_{X_2,j} \boldsymbol{N}_j + \lambda \boldsymbol{K}\right)^{-1} \sum_{j=1}^{m} \boldsymbol{N}_j' \boldsymbol{U}_{X_2,j}\, \underline{\boldsymbol{y}}_j\,, \tag{3.31}$$

and

$$\hat{\underline{\boldsymbol{\beta}}} = \left(\sum_{j=1}^{m} \boldsymbol{X}_{2j}' \boldsymbol{U}_{N,j} \boldsymbol{X}_{2,j}\right)^{-1} \sum_{j=1}^{m} \boldsymbol{X}_{2j}' \boldsymbol{U}_{N,j} \underline{\boldsymbol{y}}_j\,, \tag{3.32}$$

where

$$\boldsymbol{U}_{N,j} = \boldsymbol{V}_j^{-1} - \boldsymbol{V}_j^{-1} \boldsymbol{N}_j \Big(\sum_j \boldsymbol{N}_j' \boldsymbol{V}_j^{-1} \boldsymbol{N}_j + \lambda \boldsymbol{K}\Big)^{-1} \boldsymbol{N}_j' \boldsymbol{V}_j^{-1}\,,$$

and

$$\boldsymbol{U}_{X_2,j} = \boldsymbol{V}_j^{-1} - \boldsymbol{V}_j^{-1} \boldsymbol{X}_{2j} \Big(\sum_j \boldsymbol{X}_{2j}' \boldsymbol{V}_j^{-1} \boldsymbol{X}_{2j}\Big)^{-1} \boldsymbol{X}_{2j}' \boldsymbol{V}_j^{-1}\,,$$

with $\boldsymbol{V}_j$ given in (3.4).

The parameters $\boldsymbol{f}$ and $\boldsymbol{\beta}$ can also be estimated by the Expectation Maximization (EM) algorithm. The EM algorithm is an iterative procedure for locating the mode of the likelihood, or in Bayesian modeling for determining the posterior mode, see section 1.D. In our case, we need to maximize the penalized likelihood rather than the likelihood itself. From a Bayesian viewpoint, this does not substantially alter the problem but is merely a choice of the prior. Note that the modes of the log posterior and the penalized log-likelihood coincide if a flat prior is taken for $\boldsymbol{\beta}$, and the log-prior of $\boldsymbol{f}_{\text{css}}$ is, except for a constant, equal to $-\frac{1}{2}\lambda \boldsymbol{f}_{\text{css}}' \boldsymbol{K} \boldsymbol{f}_{\text{css}}$.

The EM algorithm consists of an E-step and an M-step. To carry out the E-step, we define the complete-data log-likelihood of $\boldsymbol{y}$ and the random coefficients $\underline{\boldsymbol{\delta}}$ (treated in this algorithm as missing data) given the model parameters, i.e. $\log\, p(\underline{\boldsymbol{y}}, \underline{\boldsymbol{\delta}} \mid \boldsymbol{f}_{\text{css}}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\xi})$. We penalize the complete-data log-likelihood with roughness penalty (3.29) and we further define the conditional distribution of missing data $\underline{\boldsymbol{\delta}}$ given $\underline{\boldsymbol{y}}$ and the model parameters, i.e., $p(\underline{\boldsymbol{\delta}} \mid \underline{\boldsymbol{y}}, \tilde{\underline{\boldsymbol{f}}}_{\text{css}}, \tilde{\underline{\boldsymbol{\beta}}}, \sigma^2, \boldsymbol{\xi})$. Here, $\boldsymbol{f}_{\text{css}}$ and $\boldsymbol{\beta}$ have been replaced by their current estimates $\tilde{\underline{\boldsymbol{f}}}_{\text{css}}$ and $\tilde{\underline{\boldsymbol{\beta}}}$. The variance components σ^2 and $\boldsymbol{\xi}$ are assumed to be known. The E-step consists of taking the expectation of the penalized complete-data log-likelihood with respect to the conditional distribution of the missing data. This involves computing the conditional expectations of $\underline{\boldsymbol{\delta}}$ and $\underline{\boldsymbol{\delta}}\,\underline{\boldsymbol{\delta}}'$, where the former expectation is the empirical Bayes estimator of the random effects.

In the M-step, we maximize the expected penalized complete-data log-likelihood (retrieved from the E-step) with respect to the model parameters $\boldsymbol{f}_{\text{css}}$ and $\boldsymbol{\beta}$. The M-step is computationally expensive if the number of distinct time points T is large because it involves inverting a $T \times T$ matrix. In that case, it is better to update the estimates of $\boldsymbol{f}_{\text{css}}$ and $\boldsymbol{\beta}$ sequentially. First, we maximize with respect to $\boldsymbol{f}_{\text{css}}$ and obtain the updated estimate

$$\tilde{\underline{\boldsymbol{f}}}_{\text{css}} = \Big(\sum_j (\boldsymbol{N}_j' \boldsymbol{N}_j) + \sigma^2 \lambda \boldsymbol{K}\Big)^{-1} \sum_j (\underline{\boldsymbol{y}}_j - \boldsymbol{X}_{2j} \tilde{\underline{\boldsymbol{\beta}}} - \boldsymbol{Z}_j \tilde{\underline{\boldsymbol{\delta}}}_j),$$

where $\tilde{\underline{\boldsymbol{\delta}}}_j$ is the empirical Bayes estimate of $\underline{\boldsymbol{\delta}}_j$ at the current estimates of $\boldsymbol{f}_{\text{css}}$ and $\boldsymbol{\beta}$. Second, we maximize with respect to $\boldsymbol{\beta}$ only and obtain the update

$$\tilde{\underline{\boldsymbol{\beta}}} = \Big(\sum_j \boldsymbol{X}_{2j}' \boldsymbol{X}_{2j}\Big)^{-1} \sum_j (\underline{\boldsymbol{y}}_j - \boldsymbol{N}_j \tilde{\underline{\boldsymbol{f}}}_{\text{css}} - \boldsymbol{Z}_j \tilde{\underline{\boldsymbol{\delta}}}_j).$$

These two steps are computationally cheap: the number of numerical operations to update the estimates of $\boldsymbol{f}_{\text{css}}$ and $\boldsymbol{\beta}$ is of order T. Although the expression for $\tilde{\underline{\boldsymbol{f}}}_{\text{css}}$ contains the inverse of a $T \times T$ matrix, efficient computation is possible using the Cholesky factorization method as described for example in Green and Silverman [21]. This algorithm where the M-step is replaced by two sequential steps is known as the EC(onditional)M algorithm [43]. The two sequential steps can also be viewed as steps of the backfitting algorithm as described by Hastie and Tibshirani [24, p. 91].

An EM algorithm can also be constructed after having reparametrized the model according to Green [20]. Using features of cubic splines, we can write $\boldsymbol{f}_{\text{css}}$ in (3.30) via a one-to-one transformation as

$$\boldsymbol{f}_{\text{css}} = \gamma_0 \mathbf{1}_T + \gamma_1 \boldsymbol{x}^* + \boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{Q})^{-1} \boldsymbol{L}\boldsymbol{\eta}, \tag{3.33}$$

where γ_0 and γ_1 are scalars, $\boldsymbol{x}^* = (x_1, \ldots, x_T)'$, $\boldsymbol{\eta}$ is a $(T-2) \times 1$ parameter vector, and $\boldsymbol{L}$ satisfies $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{R}$. For the definition of $\boldsymbol{Q}$ and $\boldsymbol{R}$, see section 3.9.

Because the columns of $\boldsymbol{Q}$ are orthogonal to $\mathbf{1}_T$ and $\boldsymbol{x}^*$, it follows that $\boldsymbol{\eta}'\boldsymbol{\eta}$ is equal to $\boldsymbol{f}'_{\text{css}}\boldsymbol{K}\boldsymbol{f}_{\text{css}}$. Hence, the penalized log-likelihood of model (3.30) with $\boldsymbol{f}_{\text{css}}$ replaced by (3.33) is equal to the sum of the log-likelihood and the term $-\frac{1}{2}\lambda\boldsymbol{\eta}'\boldsymbol{\eta}$. The E-step and M-step can be derived as before. When T is large, the computational burden can again be lowered by replacing the M-step by sequential steps.

So far, we have regarded the variance components σ^2 and $\boldsymbol{\xi}$ as known. Simple estimators of σ^2 and $\boldsymbol{\xi}$ are obtained by fitting an overelaborated model with in the fixed part T dummy predictors, one for each distinct time point [55, p. 123]. If the model with dummy effects is estimated by restricted IGLS, unbiased estimates are obtained for σ^2 and $\boldsymbol{\xi}$ also when $\underline{y}$ in the true model is associated to x by a smooth function $f(x)$. For reasons of efficiency, it may be preferable to use estimators that depend on the external smoothing parameter λ. Several authors have suggested to consider $\boldsymbol{\eta}$ as a vector of random effects $\underline{\boldsymbol{\eta}}$ and to fit a crossed random effects model with model parameters γ_0, γ_1, $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\xi}$ and random effects $\underline{\boldsymbol{\delta}}$ and $\underline{\boldsymbol{\eta}}$ [52, 58, 64]. The formulation of the crossed random effects model is attractive because it allows us to estimate $\boldsymbol{f}_{\text{css}}$ using existing software. Estimates can be obtained with the restricted IGLS algorithm implemented in MLwiN [19] and SAS [39]. Here, the variance of $\underline{\boldsymbol{\eta}}$ is set equal to the inverse of the roughness penalty λ. The restricted IGLS estimator of σ^2 performs well in simulation studies [64]. Besides, in a single level situation (e.g., longitudinal data of one subject), this estimator is equal to the classical estimate of σ^2 described for example by Green and Silverman [21, p. 39]. The estimation of the crossed random effects model via restricted IGLS is computationally demanding if the number of distinct values $x_1, \ldots, x_T$ is large because the number of crossed random effects is equal to $T - 2$.

3.9.2 Inferences

A common approach to drawing inferences about $\boldsymbol{f}_{\text{css}}$ is to construct pointwise correct confidence intervals at $x_1, \ldots, x_T$. This requires an estimate of the variance of $\hat{\underline{\boldsymbol{f}}}_{\text{css}}$. Two common estimates will be discussed. The first estimate is obtained by assuming that $\hat{\underline{\boldsymbol{f}}}_{\text{css}}$ is an estimate of the fixed, unknown $\boldsymbol{f}_{\text{css}}$. From (3.31) where $\hat{\underline{\boldsymbol{f}}}_{\text{css}}$ is written as a linear function of $\underline{\boldsymbol{y}}$, it follows that the covariance matrix is given by

$$\text{Cov}_{\text{F}}(\hat{\underline{\boldsymbol{f}}}_{\text{css}}) = \boldsymbol{W}^{-1}\left(\sum_{j=1}^{m}\boldsymbol{N}_j'\boldsymbol{U}_{X_2,j}\boldsymbol{N}_j\right)\boldsymbol{W}^{-1} \tag{3.34}$$

where

$$\boldsymbol{W} = \sum_{j=1}^{m}\boldsymbol{N}_j'\boldsymbol{U}_{X_2,j}\boldsymbol{N}_j + \lambda\boldsymbol{K}.$$

The second estimate of the variance is the posterior variance obtained from a Bayesian model where the logarithm of the prior of $\boldsymbol{f}_{\text{css}}$ is equal to $-\frac{1}{2}\lambda \boldsymbol{f}'_{\text{css}} \boldsymbol{K} \boldsymbol{f}_{\text{css}}$ except for a constant. The posterior covariance matrix has a simple form

$$\text{Cov}_{\text{B}}(\underline{\hat{\boldsymbol{f}}}_{\text{css}}) = \boldsymbol{W}^{-1}. \tag{3.35}$$

Zhang et al. [64] and Lin and Zhang [38] compare the frequentist and Bayesian estimator in a simulation study in which a fixed nonparametric function $f(x)$ is postulated. The main conclusion in these studies is that both estimators are accurate but that the Bayesian estimator sometimes performs slightly better because it accounts for the bias in $\underline{\hat{\boldsymbol{f}}}_{\text{css}}$. The Bayesian variances can also be obtained from the model with crossed random effects. Software packages such as MLwiN yield estimates of the variances of $\underline{\hat{\gamma}}_0$ and $\underline{\hat{\gamma}}_1$ and the comparative variance of the empirical Bayes estimator $\underline{\hat{\boldsymbol{\eta}}}$. The covariance between $\underline{\hat{\boldsymbol{\eta}}}$ and $(\underline{\hat{\gamma}}_0, \underline{\hat{\gamma}}_1)$ is not always produced. However, the design matrices of (γ_0, γ_1) and $\underline{\boldsymbol{\eta}}$ are orthogonal if the points at which the measurements are taken are common to all subjects. Therefore, the precision of the estimator of $\text{Cov}_{\text{B}}\big(\underline{\hat{\boldsymbol{f}}}_{\text{css}}\big)$ is in general not substantially affected by the omission of the covariance between $(\underline{\hat{\gamma}}_0, \underline{\hat{\gamma}}_1)$ and $\underline{\hat{\boldsymbol{\eta}}}$.

Besides the Bayesian model with a finite-dimensional prior for $\boldsymbol{f}_{\text{css}}$, a model with an infinite-dimensional prior for the continuous spline $f_{\text{css}}(x)$ exists as well [58, 64]. This model was put forward by Wahba [56] and is appealing because a smoothing spline $f_{\text{css}}(x)$ is defined for all x and not only for the observed values. The finite- and infinite-dimensional formulation lead to the same posterior variance of $\underline{\hat{\boldsymbol{f}}}_{\text{css}}$.

3.9.3 Smoothing Parameter Selection

Several methods exist for selecting the smoothing parameter λ. In this section, three are discussed. The first method is to maximize the cross-validated log-likelihood as a function of λ. The cross-validated log-likelihood is an approximation to the expectation of the predictive log-likelihood which is the expected log-likelihood of a new vector of observations $\underline{\boldsymbol{y}}^*$ at the penalized likelihood estimators of the model parameters $\boldsymbol{f}_{\text{css}}$, $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\xi}$. The prediction process is imitated by leaving out one subject at a time and predicting the omitted subject on the basis of the other subjects' data [46].

A drawback of cross-validation is that it is computationally expensive. An alternative strategy is to estimate the expected predictive log-likelihood by the sum of the log-likelihood and the trace of the matrix $\boldsymbol{A}$ that maps $\underline{\boldsymbol{y}}$ on the estimator $\underline{\hat{\boldsymbol{f}}}_{\text{css}} = \boldsymbol{A}\underline{\boldsymbol{y}}$ [24, p. 52; 21, p. 37]. This estimator, named Mallows' C_p, is cheap and unbiased if the (co)variance parameters σ^2 and $\boldsymbol{\xi}$ are known. For uncorrelated data, the unbiasedness proof is provided by Hastie and Tibshirani [24, p. 48]. The proof in the case of multilevel data

is analogous. In practice, σ^2 and $\boldsymbol{\xi}$ are unknown and can be estimated by the restricted IGLS estimators in the overelaborated model with dummy time predictors (see section 3.9.1).

A limitation of applying criteria like the cross-validated log-likelihood or Mallows' C_p is that λ is not treated as a model parameter but as an external variable. The smoothing parameter becomes a model parameter if we adopt the crossed random effects model and estimate the variance of $\underline{\boldsymbol{\eta}}$ freely instead of constraining the variance to be equal to the inverse of λ. It can be shown that if the crossed random effects model is estimated by restricted IGLS implemented in MLwiN [19], then the estimate of λ is the generalized maximum likelihood (GML) estimate [57, 64] which has good performance in simulation studies [33]. It may also be sensible to examine whether a model with smoothing spline $f_{\text{css}}(x)$ fits better than a model with a linear effect for x. Hastie and Tibshirani [25, p. 65] provide some approximate F-tests based on residual sums of squares and Cantoni and Hastie [9] and Guo [22] present likelihood ratio tests for $H_0 : \lambda^{-1} = 0$ which is equivalent to $H_0 : \underline{\boldsymbol{\eta}} = 0$ (3.33). Instead of the likelihood ratio test, the score test may also be considered. The score test is computationally cheap because estimates of the model with crossed random effects are not required. The test is based on the one-step estimator that is obtained when we start from the estimate of the null model. The ratio of the one-step estimator to its standard error has an asymptotic standard normal null distribution. The score test also has good power properties in a small sample setting [6]. For testing against an unspecified but monotonic effect of x, this test against a linear effect may be expected to have good power against most non-linear effects.

3.10 Example: Effect of IQ on a Language Test

We fitted the three different functions that were discussed so far, i.e., the polynomial function, the regression spline function, and the cubic smoothing spline function, to a real data set. The estimations were done using MLwiN 1.1 [19] and Gauss 3.2 [1]. The data set is described in Snijders and Bosker [51]. It contains language test scores of 2287 pupils within 131 elementary schools. We modeled the test score (Y) as a function of the grand-mean centered IQ of the pupil (`IQ`), the gender of the pupil (`SEX`), the school average of IQ ($\overline{\texttt{IQ}}$), and the socio-economic status (`SES`) of the pupil. We assumed a non-linear effect for `IQ` and linear effects for the other predictors. Note that in most applications of models with functional non-linear effects, time is the ordering principle but an ordering according to any other unidimensional variable is possible as well. Between-school differences were modeled by including a random intercept and a random slope of `IQ` at level two. Finally, we assumed that the level-one measurement errors are homoscedastic and uncorrelated. The model can be

written as

$$\underline{y}_{ij} = f(\texttt{IQ}_{ij}) + \beta_1 \texttt{SES}_{ij} + \beta_2 \texttt{SEX}_{ij} + \beta_3 \overline{\texttt{IQ}}_j + \underline{\delta}_{0j} + \underline{\delta}_{1j}\texttt{IQ}_{ij} + \underline{\epsilon}_{ij} \,.$$

The estimated polynomial function, regression spline function, and cubic smoothing spline function are presented in Figure 3.3. The chosen polynomial function is of order three. We also considered a fourth-degree polynomial but this did not yield a further improvement in fit. The chosen regression spline is a quadratic spline with a knot at zero. This function was considered by Snijders and Bosker [51, p. 113] as a flexible and parsimonious alternative for the polynomial function. We determined the smoothness of the cubic smoothing spline by maximizing GML. We also considered optimization of the cross-validated log-likelihood and Mallows' C_p but the three methods rendered similar values for the smoothing parameter: $\lambda_{\mathrm{GML}} = 1.6$, $\lambda_{C_p} = 1.6$, $\lambda_{\mathrm{CV}} = 2.0$.

The three fitted functions lead to similar predictions: the effect of IQ on Y is larger in the middle than in the tails of the distribution of IQ. The smoothing spline performs slightly better than the other two functions since it is monotonically increasing whereas the polynomial function and the regression spline have a negative slope at low and high values of IQ.

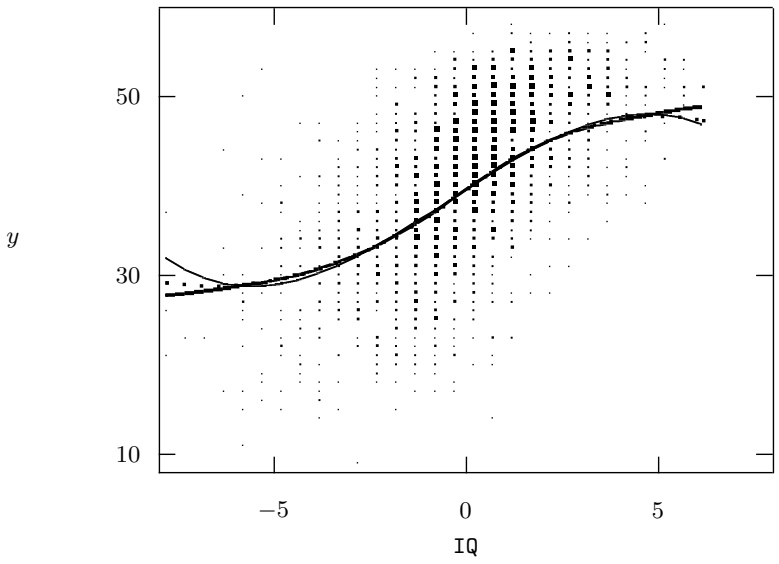

Fig. 3.3. Language test score (y) against centered IQ score: raw data, cubic polynomial estimate (thin), quadratic regression spline estimate (dashed), and smoothing spline estimate (bold).

We also estimated the pointwise standard errors of the fitted functions. These are presented in Figure 3.4. We see that the standard errors of the fitted functions are very similar. Data are sparse at the left and right end of the window (Figure 3.4) and the standard errors are large there compared to the middle part. We further see that the Bayesian standard error of the cubic smoothing spline estimate is slightly larger than its frequentist counterpart, as it should be according to (3.34) and (3.35) [cf. 64].

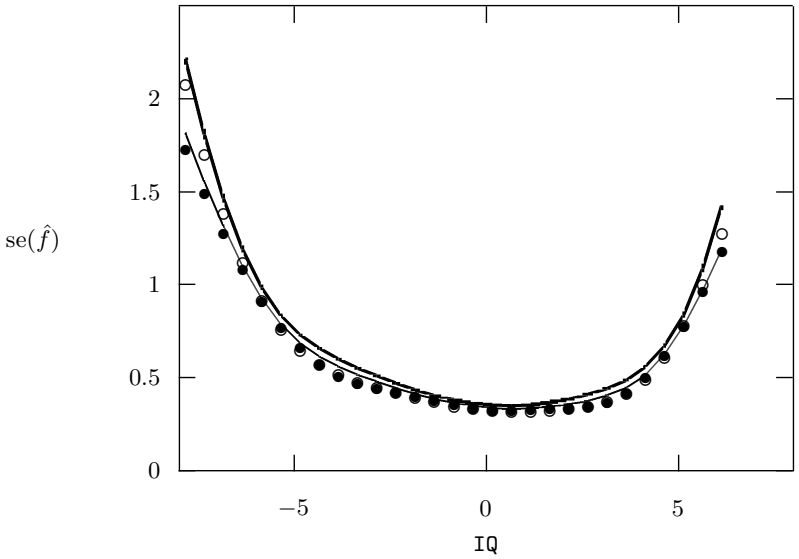

Fig. 3.4. Standard errors of the cubic polynomial estimate (open circle) and the quadratic regression spline estimate (closed circle), and Bayesian (bold line) and frequentist (thin line) standard errors of the smoothing spline estimate.

3.11 Extensions

The model can be extended to a model with more than two levels or a model with non-normal responses in the same way as multilevel models without a functional effect can be extended. Another direction is to specify a model with two functional effects, $f(x)$ and $g(v)$. This model is called an additive model and is put forward by Hastie and Tibshirani [24]. Algorithms for estimating additive multilevel models with cubic smoothing splines are provided by Lin and Zhang [38]. A related model is a model in which the effect of predictor w on y is described by function $h(x) \times w$. This model is known as the varying coefficient model and has been used to describe time-varying effects of

predictors in longitudinal studies [25]. A multilevel extension of the model is presented by Hoover et al. [29]. The additive and varying coefficient models can be formulated as random effects models with a separate random effect for each functional effect. The estimation can be done in MLwiN but becomes demanding if we have many functional effects. For varying coefficient models, less demanding estimators are available [10, 15].

We have discussed functional effects to describe the mean pattern. Functional effects for the random part of the model have been proposed as well. In multilevel modeling, a common, simple choice is to include polynomial functions in the random part of the model [cf. 8, 17, 50]. When adding spline functions instead of polynomial functions to the random part, a possible approach is to define a separate smoothing spline for each level-two unit and to use the mixed effects formulation to define a nested sample of curves [7, 22]. The mixed effects approach is appealing, but it is computationally demanding when the number of distinct points is large. A somewhat different approach is to explore the covariance structure by a principal components analysis yielding functions that describe the main sources of variation among the individual curves. These methods are particularly attractive when studying variability between individual curves. Rice and Silverman [46] propose a principal components model where the differences among individuals are described by cubic smoothing splines. The model is applicable only when the points at which measurements are taken are common to all level-two units. Rice and Wu [47] and James et al. [31] use B-spline functions to allow for irregular spacing of the data. Yao et al. [62] present a model for irregular data with functions retrieved from a smooth estimate of the (continuous) covariance surface. For the underlying functions, they also provide asymptotic confidence bounds.

Acknowledgement. The research by Johannes Berkhof was financially supported by grant ESR 510-78-501 of the Dutch Organization for Scientific Research (NWO).

References

1. Aptech Systems. *Gauss*. Aptech Systems, Maple Valley, WA, 1994.
2. A. C. Atkinson. *Plots, Transformations, and Regression*. Clarendon Press, Oxford, UK, 1985.
3. B. H. Baltagi. *Econometric Analysis of Panel Data*. Wiley, New York, 1995.
4. M. S. Bartlett and D. G. Kendall. The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, 8:128–138, 1946.
5. C. S. Berkey and R. L. Kent Jr. Longitudinal principal components and nonlinear regression models of early childhood growth. *Annals of Human Biology*, 10:523–536, 1983.

6. J. Berkhof and T. A. B. Snijders. Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, 26:133–152, 2002.
7. B. A. Brumback and J. A. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93:961–994, 1998. (with discussion)
8. A. S. Bryk and S. W. Raudenbush. Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101:147–158, 1987.
9. E. Cantoni and T. Hastie. Degrees-of-freedom tests for smoothing splines. *Biometrika*, 89:251–263, 2002.
10. C.-T. Chiang, J. A. Rice, and C. O. Wu. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96:605–619, 2001.
11. R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman & Hall, London, 1982.
12. C. de Boor. *A Practical Guide to Splines*. Springer, New York, 1978.
13. L. E. Eberly and L. M. Thackeray. On Lange and Ryan's plotting technique for diagnosing non-normality of random effects. *Statistics & Probability Letters*, 75: 77–85, 2005.
14. P. H. C. Eilers and B. D. Marx. Flexible smoothing with splines and penalties. *Statistical Science*, 11:89–121, 1996. (with discussion)
15. J. Fan and J. T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B*, 62:303–322, 2000.
16. A. Fielding. Role of the Hausman test and whether higher level effects should be treated as random or fixed. *Multilevel Modelling Newsletter*, 16(2):3–9, 2004.
17. H. Goldstein. Efficient statistical modelling of longitudinal data. *Annals of Human Biology*, 13:129–141, 1986.
18. H. Goldstein. *Multilevel Statistical Models*. Edward Arnold, London, 3rd edition, 2003.
19. H. Goldstein, J. Rasbash, I. Plewis, D. Draper, W. Browne, M. Yang, G. Woodhouse, and M. Healy. *A User's Guide to MLwiN*. Multilevel Models Project, Institute of Education, University of London, London, 1998.
20. P. J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55:245–259, 1987.
21. P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London, 1994.
22. W. Guo. Functional mixed effects model. *Biometrics*, 58:121–128, 2002.
23. J. Haslett and D. Dillane. Application of 'delete = replace' to deletion diagnostics for variance component estimation in the linear mixed model. *Journal of the Royal Statistical Society, Series B*, 66:131–143, 2004.
24. T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall, London, 1990.
25. T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55:757–796, 1993. (with discussion)
26. J. A. Hausman. Specification tests in econometrics. *Econometrica*, 46:1251–1271, 1978.

27. J. A. Hilden-Minton. *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*. PhD thesis, Department of Mathematics, University of California, Los Angeles, 1995.
28. J. S. Hodges. Some algebra and geometry for hierarchical linear models, applied to diagnostics. *Journal of the Royal Statistical Society, Series B*, 60:497–536, 1998.
29. D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85:809–822, 1998.
30. C. Hsiao. Random coefficient models. In L. Mátyás and P. Sevestre, editors, *The Econometrics of Panel Data*, pages 77–99. Kluwer, Dordrecht, The Netherlands, 2nd edition, 1996.
31. G. James, T. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000.
32. J.-S. Kim and E. W. Frees. Omitted variables in multilevel models. *Psychometrika*, in press.
33. R. Kohn, C. F. Ansley, and D. Tharm. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, 86:1042–1050, 1991.
34. N. Lange and L. Ryan. Assessing normality in random effects models. *Annals of Statistics*, 17:624–642, 1989.
35. I. H. Langford and T. Lewis. Outliers in multilevel data. *Journal of the Royal Statistical Society, Series A*, 161:121–160, 1998.
36. E. Lesaffre and G. Verbeke. Local influence in linear mixed models. *Biometrics*, 54:570–582, 1998.
37. T. Lewis and I. H. Langford. Outliers, robustness and the detection of discrepant data. In A. H. Leyland and H. Goldstein, editors, *Multilevel Modelling of Health Statistics*, pages 75–91. Wiley, New York, 2001.
38. X. Lin and D. Zhang. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 61:381–400, 1999.
39. R. C. Littell, G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. *SAS System for Mixed Models*. SAS Institute, Cary, NC, 1996.
40. N. T. Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74:817–827, 1987.
41. N. T. Longford. *Random Coefficient Models*. Oxford University Press, Oxford, UK, 1993.
42. N. T. Longford. Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society, Series A*, 164:259–273, 2001.
43. X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278, 1993.
44. C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 2nd edition, 1973.
45. S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Thousand Oaks, CA, 2nd edition, 2002.

46. J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, 53:233–243, 1991.
47. J. A. Rice and C. O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57:253–259, 2001.
48. M. H. Seltzer, W. H. Wong, and A. S. Bryk. Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics*, 21:131–167, 1996.
49. A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL, 2004.
50. T. A. B. Snijders. Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, 30:405–426, 1996.
51. T. A. B. Snijders and R. J. Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage, London, 1999.
52. T. P. Speed. Comment on "That BLUP is a good thing: the estimation of random effects" (by G. K. Robinson). *Statistical Science*, 6:44, 1991.
53. R. Van der Leeden, E. Meijer, and F. M. T. A. Busing. Resampling multilevel models. In J. de Leeuw and E. Meijer, editors, *Handbook of Multilevel Analysis*, chapter 11. Springer, New York, 2007. (this volume)
54. G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91:217–221, 1996.
55. G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000.
56. G. Wahba. Bayesian "confidence" intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, 45:133–150, 1983.
57. G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 4: 1378–1402, 1985.
58. Y. Wang. Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, 60:159–174, 1998.
59. C. Waternaux, N. M. Laird, and J. H. Ware. Methods for analysis of longitudinal data: Blood lead concentrations and cognitive development. *Journal of the American Statistical Association*, 84:33–41, 1989.
60. S. Weisberg. *Applied Linear Regression*. Wiley, New York, 3rd edition, 2005.
61. S. Wold. Spline functions in data analysis. *Technometrics*, 16:1–11, 1974.
62. F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005.
63. T. Zewotir and J. S. Galpin. Influence diagnostics for linear mixed models. *Journal of Data Science*, 3:153–177, 2005.
64. D. Zhang, X. Lin, J. Raz, and M. Sowers. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93:710–719, 1998.