

Chapter 4

DIAGNOSTIC CHECKS FOR MULTILEVEL MODELS

Tom A.B. Snijders, ICS, University of Groningen
Johannes Berkhof, Free University, Amsterdam

To be published in Jan de Leeuw & Ita Kreft (eds.),
Handbook of Quantitative Multilevel Analysis.

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

$$\underline{\mathbf{y}}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{Z}_j \underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j, \quad j = 1, \dots, m, \quad (4.1a)$$

with

$$\begin{bmatrix} \underline{\boldsymbol{\epsilon}}_j \\ \underline{\boldsymbol{\delta}}_j \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\emptyset} \\ \boldsymbol{\emptyset} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_j(\boldsymbol{\theta}) & \boldsymbol{\emptyset} \\ \boldsymbol{\emptyset} & \boldsymbol{\Omega}(\boldsymbol{\xi}) \end{bmatrix} \right) \quad (4.1b)$$

and

$$(\underline{\boldsymbol{\epsilon}}_j, \underline{\boldsymbol{\delta}}_j) \perp (\underline{\boldsymbol{\epsilon}}_\ell, \underline{\boldsymbol{\delta}}_\ell) \quad (4.1c)$$

for all $j \neq \ell$. The lengths of the vectors \mathbf{y}_j , $\boldsymbol{\beta}$, and $\boldsymbol{\delta}_j$, respectively, are n_j , r , and s . Like in all regression-type models, the explanatory variables \mathbf{X} and \mathbf{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 \mathbf{X} and \mathbf{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.

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

$$\boldsymbol{\Sigma}_j(\boldsymbol{\theta}) = \sigma^2 \mathbf{I}_{n_j}, \quad (4.1d)$$

where \mathbf{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.

Questioning this model specification can be aimed at various aspects: the choice of variables included in \mathbf{X} , the choice of variables for \mathbf{Z} , the residuals having expected value 0, the homogeneity of the covariance matrices across groups, the specification of the covariance matrices, and the multivariate normal distributions. These different aspects of the model specification are intertwined, 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 a significant random slope variance, which then disappears if the heteroscedasticity is taken into account; non-linear effects of some variables in \mathbf{X} , when unrecognized, may show up as heteroscedasticity at level 1 or as a random slope; and non-zero expected residuals sometimes can be dealt with by transformations of variables in \mathbf{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.

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, which often is 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 OLS regression model. This section treats some examples of model specification checks which do not have direct parallels in the OLS regression model.

2.1 Heteroscedasticity

The comprehensive nature of most algorithms for estimating the HLM makes it relatively straightforward to include some possibilities for modeling heteroscedasticity, 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* [Goldstein, 1995; Goldstein et al., 1998] accomodates variances depending as linear or quadratic functions of variables. For level-one

heteroscedasticity, this is carried out formally by writing

$$\underline{\epsilon}_{ij} = \mathbf{v}_{ij} \underline{\epsilon}_{ij}^0 \quad (4.2a)$$

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

$$\underline{\epsilon}_{ij}^0 \sim \mathcal{N}(0, \Sigma^0(\boldsymbol{\theta})). \quad (4.2b)$$

This implies

$$\text{Var } \underline{\epsilon}_{ij} = \mathbf{v}_{ij} \Sigma^0(\boldsymbol{\theta}) \mathbf{v}'_{ij}. \quad (4.3)$$

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

The IGLS algorithm works only with the expected values and covariance matrices implied by the model specification, see Goldstein [1995], p. 38–40. A sufficient condition for model (4.1a-4.1c) to be a meaningful representation is that (4.3) is nonnegative for all i, j – clearly less restrictive than Σ^0 being positive definite. Therefore it is not required that Σ^0 be positive definite, but it is sufficient that (4.3) is positive for all observed \mathbf{v}_{ij} . E.g., a level-one variance function depending linearly on \mathbf{v} is obtained by defining

$$\Sigma^0(\boldsymbol{\theta}) = (\sigma_{hk}(\boldsymbol{\theta}))_{1 \leq h, k \leq t} \quad (4.4a)$$

with

$$\begin{aligned} \sigma_{h1}(\boldsymbol{\theta}) &= \sigma_{1h}(\boldsymbol{\theta}) = \theta_h & h = 1, \dots, t \\ \sigma_{hk}(\boldsymbol{\theta}) &= 0 & \min\{h, k\} \geq 2 \end{aligned} \quad (4.4b)$$

where $\boldsymbol{\theta}$ is a $t \times 1$ vector. Quadratic variance functions can be represented by letting Σ^0 be a symmetric matrix, subject only to a positivity restriction for (4.3).

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 \mathbf{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 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 [1995], Chapter 3, and Snijders and Bosker [1999], Chapter 8.

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 $\underline{\delta}_j$ in (4.1) as random variables, or should they rather be treated as fixed parameters δ_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 conceptually meaningful anyway) and the statistical inference is directed at this population, then a random coefficient model is in principle appropriate. If 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. If the differences between the level-two units are a nuisance factor rather than a point of independent interest, the analysis could be done in principle either way.

On a more practical level, the choice between random and fixed effects depends also on 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 (4.1) for the random effects $\underline{\delta}_j$ 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 [1996]) or mixtures of normal distributions (the heterogeneity model of Verbeke and Lesaffre [1996], also see Verbeke and Molenberghs [2000]). 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 means are zero, we first focus on the simplest case of a random intercept model, where \mathbf{Z}_j contains only the constant vector with all its n_j entries equal to 1, expressed as $\mathbf{Z}_j = \mathbf{1}_j$. Subsequently we shall give a more formal treatment of a more general case.

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

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

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

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

$$\mathbf{X}_j = \bar{\mathbf{X}}_j + \tilde{\mathbf{X}}_j \quad (4.5a)$$

where

$$\bar{\mathbf{X}}_j = \mathbf{1}_j(\mathbf{1}'_j\mathbf{1}_j)^{-1}\mathbf{1}'_j\mathbf{X}_j. \quad (4.5b)$$

Then the data-generating model can be written as

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

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

A bias in the estimation of β will be caused by lack of orthogonality of the matrices

$$\mathbf{X}_j = \bar{\mathbf{X}}_j + \tilde{\mathbf{X}}_j \text{ and } \mathbf{1}_j\mathbf{z}_{2j}.$$

Since the definition of $\tilde{\mathbf{X}}_j$ implies that $\tilde{\mathbf{X}}_j$ is orthogonal to $\mathbf{1}_j\mathbf{z}_{2j}$, it is clear that $\bar{\mathbf{X}}_j$ is the villain of the piece: analogous to the situation of a misspecified OLS regression model, there will be a bias if $\bar{\mathbf{X}}_j'\mathbf{1}_j\mathbf{z}_{2j} \neq \mathbf{0}$. In words, this means that the level-two variables \mathbf{z}_2 on which the expected value of the random effects $\underline{\delta}_j$ depends, are not orthogonal to the group means collected in $\bar{\mathbf{X}}$. If there is such a non-orthogonality, there is an obvious solution: extend the fixed part by giving separate fixed parameters β_1 to the group means $\bar{\mathbf{X}}$ and β_2 to the deviation variables $\tilde{\mathbf{X}}$, so that the working model reads

$$\underline{\mathbf{y}}_j = \bar{\mathbf{X}}_j\beta_1 + \tilde{\mathbf{X}}_j\beta_2 + \mathbf{1}_j\tilde{\delta}_j + \underline{\epsilon}_j$$

(taking out the zero columns from $\bar{\mathbf{X}}_j$ and $\tilde{\mathbf{X}}_j$, which are generated by columns in \mathbf{X}_j which themselves are within-cluster deviation variables or level-two

variables, respectively). An equivalent working model is obtained by adding to (4.1) the fixed effects of the non-constant cluster means \bar{X}_j . In this way, the bias in the fixed effect estimates due to ignoring the term $z_{2j}\gamma$ is absorbed completely by the parameter estimate for β_1 , and this misspecification does not affect the unbiasedness of the estimate for β_2 . The estimate for the level-2 variance $Var(\underline{\delta}_j)$ will be affected, which is inescapable if there is no knowledge about 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 group means often have a substantively meaningful interpretation, different from the level-one variables from which they are calculated (cf. Snijders and Bosker [1999]). This often is a substance-matter related rationale for including the group 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 \mathbf{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-group regression coefficient and the between-group coefficient. In economics this test – or at least a test with the same purpose – is often referred to as the Hausman test. (Hausman [1978] proposed a general procedure for tests of model specification, of which the test for equality of the within-group and between-group coefficients is an important special case. Also see Baltagi [1995], who shows on p. 69 that this case of the Hausman test is equivalent to testing the effect of the cluster means \bar{X} .)

In econometrics, the Hausman test for the difference between the within-group and between-group regression coefficients is often seen as a test for deciding whether to use a random or fixed coefficient model for the level-two residuals δ_j . The preceding discussion shows that this is slightly beside the point. If there is a difference between the within-group and between-group regression coefficients, which is what this Hausman test intends to detect, then unbiased estimates for the fixed within-group effects can be obtained also with random coefficient models, provided that the group means of the explanatory variables are included among the fixed effect variables \mathbf{X} . Including the group 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.

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

model be given by

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

instead of the assumed value of 0. It may be assumed that $\mathbf{Z}_j\mathbf{Z}_{2j}$ cannot be expressed as a linear combination $\mathbf{X}_j\mathbf{A}$ for some matrix \mathbf{A} independent of j , because otherwise the contribution caused by $E\tilde{\boldsymbol{\delta}}_j$ could be absorbed into $\mathbf{X}_j\boldsymbol{\beta}$.

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

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

and the difference variables

$$\tilde{\mathbf{X}}_j = \mathbf{X}_j - \vec{\mathbf{X}}_j \text{ and } \tilde{\mathbf{y}}_j = \mathbf{y}_j - \vec{\mathbf{y}}_j .$$

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

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

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

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

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

where

$$\mathbf{V}_j = \sigma^2\mathbf{I}_{n_j} + \mathbf{Z}_j\boldsymbol{\Omega}(\boldsymbol{\xi})\mathbf{Z}'_j . \quad (4.6)$$

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

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

The inverse of \mathbf{V}_j can be written as [Rao, 1973; Longford, 1993]

$$\mathbf{V}_j^{-1} = \sigma^{-2}\mathbf{I}_{n_j} - \mathbf{Z}_j\mathbf{A}_j\mathbf{Z}'_j , \quad (4.7)$$

for a matrix

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

This implies that

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

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

$$\begin{aligned} & -\frac{1}{2} \sum_j \left(\log \det(\mathbf{V}_j) \right. \\ & \quad + (\vec{\mathbf{y}}_j - \vec{\mathbf{X}}_j\boldsymbol{\beta} - \mathbf{Z}_j\mathbf{Z}_{2j}\boldsymbol{\gamma})' \mathbf{V}_j^{-1} (\vec{\mathbf{y}}_j - \vec{\mathbf{X}}_j\boldsymbol{\beta} - \mathbf{Z}_j\mathbf{Z}_{2j}\boldsymbol{\gamma}) \\ & \quad \left. + \sigma^{-2} \|\tilde{\mathbf{y}}_j - \tilde{\mathbf{X}}_j\boldsymbol{\beta}\|^2 \right). \end{aligned} \quad (4.8)$$

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

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

the bias due to neglecting the term $\mathbf{Z}_{2j}\boldsymbol{\gamma}$ in the expected value of $\boldsymbol{\delta}_j$ will be absorbed into the estimate of $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ will be an unbiased estimate for the fixed effect of \mathbf{X} . The log-likelihood (4.8) shows that the ML and REML estimates of $\boldsymbol{\beta}_2$ are equal to the OLS estimate based on the deviation variables $\vec{\mathbf{y}}_j$, and also equal to the OLS estimate in the model obtained by replacing the random effects $\boldsymbol{\delta}_j$ 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., m_s) of fixed effects, it is advisable to add to the model the fixed effects of the variables

$$\vec{\mathbf{X}}_j = \mathbf{Z}_j(\mathbf{Z}'_j\mathbf{Z}_j)^{-1}\mathbf{Z}'_j\mathbf{X}_j, \quad (4.9)$$

i.e., the predictions of the variables in \mathbf{X} by within-cluster OLS regression of \mathbf{X}_j on \mathbf{Z}_j . The model term $\mathbf{Z}_j\boldsymbol{\delta}_j$ will be entirely absorbed into the fixed effects of $\vec{\mathbf{X}}_j$, and the estimates of $\boldsymbol{\beta}_2$ will be unbiased for the corresponding elements of $\boldsymbol{\beta}$ in (4.1). Depending on the substantive context, there may well be a meaningful interpretation of the constructed level-two variables (4.9).

3. Residuals

Like in other regression-type models, residuals 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. E.g., Bryk and Raudenbush [1992] suggest an upward approach for model construction, whereas Langford and Lewis [1998] propose a downward approach for the purpose of outlier inspection. In our view, the argument given by Hilden-Minton [1995] 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 3.5.

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 [1995], as the OLS residuals calculated separately within each level-two cluster. These will be called here the OLS within-cluster residuals. Consider again model (4.1) with the further specification (4.1d). When $\tilde{\mathbf{X}}_j$ is the matrix containing all non-redundant columns in $(\mathbf{X}_j \mathbf{Z}_j)$ and \mathbf{P}_j is the corresponding projection matrix (the "Hat matrix")

$$\mathbf{P}_j = \tilde{\mathbf{X}}_j (\tilde{\mathbf{X}}_j' \tilde{\mathbf{X}}_j)^{-1} \tilde{\mathbf{X}}_j',$$

the OLS within-cluster residuals are given by

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

The model definition implies that

$$\hat{\boldsymbol{\epsilon}}_j = (\mathbf{I}_{n_j} - \mathbf{P}_j) \boldsymbol{\epsilon}_j, \quad (4.10)$$

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

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 \mathbf{X} and \mathbf{Z} . Linearity of the dependence on these variables can be checked by plotting the residuals $\hat{\boldsymbol{\epsilon}}_j$ against the variables in \mathbf{X} and \mathbf{Z} . The presence of outliers and potential effects of omitted but available variables can be studied analogously.

In the second place, the homoscedasticity assumption (4.1d) can be checked. Equation (4.10) implies that, if the model assumptions are correct,

$$\hat{\underline{\epsilon}}_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2(1 - h_{ij})) \quad (4.11)$$

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

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

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 (e.g., Green and Silverman [1994]), choosing the smoothing parameter so as to minimize the cross-validated 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 [1985] has an illuminating discussion of non-linear transformations of the dependent variable in single-level regression models. Hodges [1998], p. 506, discusses Box-Cox transformations for multi-level models.

Example 3.1. As an example, consider the data set provided with the *MLwiN* software [Goldstein et al., 1998] 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-1 residuals will be illustrated.

When the OLS within-cluster residuals are plotted against the explanatory variable `standlrt`, an unilluminating cloud of points is produced. Therefore only the smoothing spline approximating these residuals is plotted in Figure 4.1.

	Model 1		Model 2		Model 3	
<i>Fixed part</i>						
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)
<i>Random part</i>						
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: σ_1^2					-0.014	(.006)
deviance	9357.2		9346.2		9341.4	

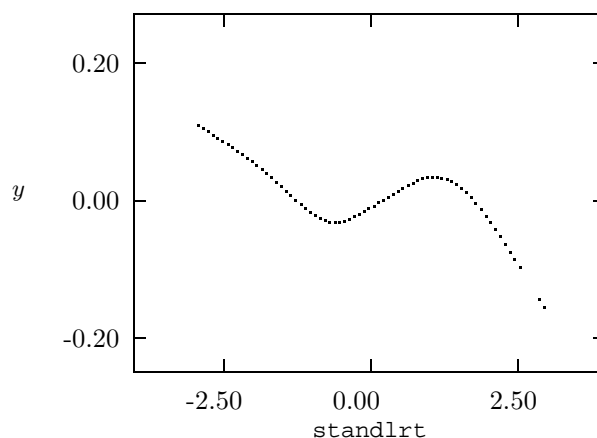


Figure 4.1. Smoothing spline approximation for OLS within-cluster residuals (y) under Model 1 against standardized reading test (standlrt).

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 (4.11) are calculated for Model 2. The smoothing spline approximating the squared semi-standardized residuals is plotted against standlrt in Figure 4.2.

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

$$\text{Var}(\epsilon_{ij}) = \sigma^2 + \sigma_1^2 \text{standlrt}_{ij},$$

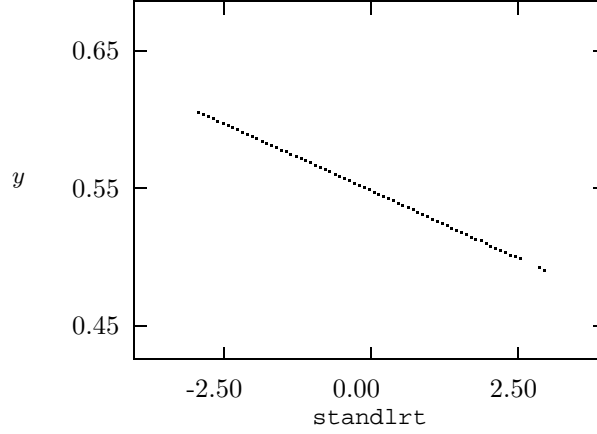


Figure 4.2. Smoothing spline approximation for squared semi-standardized OLS within-cluster residuals (y) under Model 2 against standardized reading test (standlrt).

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

3.2 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 (4.1d) is tested against the alternative

$$\Sigma_j(\theta) = \sigma_j^2 \mathbf{I}_{n_j},$$

where the σ_j^2 are unspecified and not identical.

Indicating the rank of $\tilde{\mathbf{X}}_j$ defined in Section 3.1 by r_j , the within-cluster residual variance is

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

If model (4.1d) is correct, $(n_j - r_j)\underline{s}_j^2/\sigma^2$ has a chi-squared distribution on $(n_j - r_j)$ degrees of freedom. The homogeneity test of Bartlett and Kendall [1946] can be applied here (it is proposed also in Bryk and Raudenbush [1992], p. 208, and Snijders and Bosker [1999], p. 127). Denoting $\sum n_j = n_+$, $\sum r_j = r_+$ and

$$l_{\underline{s}_{\text{pooled}}} = \frac{1}{n_+ - r_+} \sum_j (n_j - r_j) \ln(\underline{s}_j^2), \quad (4.12)$$

the test statistic is given by

$$\underline{H} = \sum_j \frac{n_j - r_j}{2} (\ln(\underline{s}_j^2) - \underline{l}_{s_{\text{pooled}}})^2. \quad (4.13)$$

Under the null hypothesis this statistic has approximately a chi-squared distribution on $\tilde{m} - 1$ 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 (4.13) 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 on $(n_j - r_j)$ d.f. and applying formulae (4.12, 4.13) 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 Level-two residuals

There are two main ways for predicting¹ the level-two residuals $\underline{\delta}_j$: the OLS method (based on treating them as fixed effects δ_j) and the empirical Bayes (EB) method. The empirical Bayes ‘estimate’ of $\underline{\delta}_j$ can be defined as its conditional expected value given the observations $\underline{y}_1, \dots, \underline{y}_m$, plugging in the parameter estimates for β, θ , and ξ . (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. [1989] and Hilden-Minton [1995]. Their conclusion was that, provided the level-one model (i.e., the assumptions about the level-one predictors included in \mathbf{X} and about the level-one residuals $\underline{\epsilon}_j$) is adequate, it is advisable to use the EB estimates.

¹Traditional 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 consequently respect this terminology, since almost everybody writes about *estimation* of residuals.

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

$$\begin{aligned}\hat{\underline{\boldsymbol{\delta}}}_j &= E\{\underline{\boldsymbol{\delta}}_j \mid \underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_m\} \text{ (using parameter estimates } \hat{\underline{\boldsymbol{\beta}}}, \hat{\underline{\boldsymbol{\theta}}}, \hat{\underline{\boldsymbol{\xi}}}\text{)} \\ &= \hat{\underline{\boldsymbol{\Omega}}}\mathbf{Z}'_j\hat{\underline{\mathbf{V}}}_j^{-1} \left(\underline{\mathbf{y}}_j - \mathbf{X}_j\hat{\underline{\boldsymbol{\beta}}}_j \right) \\ &= \hat{\underline{\boldsymbol{\Omega}}}\mathbf{Z}'_j\hat{\underline{\mathbf{V}}}_j^{-1} \left(\mathbf{Z}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j - \mathbf{X}_j(\hat{\underline{\boldsymbol{\beta}}} - \boldsymbol{\beta}) \right)\end{aligned}\quad (4.14)$$

where

$$\mathbf{V}_j = \text{Cov}(\underline{\mathbf{y}}_j) = \mathbf{Z}_j\boldsymbol{\Omega}\mathbf{Z}'_j + \boldsymbol{\Sigma}_j, \quad (4.15)$$

$$\hat{\underline{\mathbf{V}}}_j = \mathbf{Z}_j\hat{\underline{\boldsymbol{\Omega}}}\mathbf{Z}'_j + \hat{\underline{\boldsymbol{\Sigma}}}_j, \quad (4.16)$$

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

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

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

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 Bryk and Raudenbush [1992], p. 43.

The EB residuals can be expressed as

$$\begin{aligned}\hat{\underline{\boldsymbol{\delta}}}_j &= \hat{\underline{\mathbf{R}}}_j\underline{\boldsymbol{\delta}}_j + \hat{\underline{\boldsymbol{\Omega}}}\mathbf{Z}'_j\hat{\underline{\mathbf{V}}}_j^{-1}\underline{\boldsymbol{\epsilon}}_j \\ &\quad - \hat{\underline{\boldsymbol{\Omega}}}\mathbf{Z}'_j\hat{\underline{\mathbf{V}}}_j^{-1}\mathbf{X}_j(\hat{\underline{\boldsymbol{\beta}}} - \boldsymbol{\beta}).\end{aligned}\quad (4.17)$$

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 $\boldsymbol{\beta}$.

The covariance matrix of the EB residuals is

$$\text{Cov}(\hat{\underline{\boldsymbol{\delta}}}_j) = \boldsymbol{\Omega}\mathbf{Z}'_j\mathbf{V}_j^{-1} \left(\mathbf{V}_j - \mathbf{X}_j \left(\sum_{\ell=1}^m \mathbf{X}'_{\ell}\mathbf{V}_{\ell}^{-1}\mathbf{X}_{\ell} \right)^{-1} \mathbf{X}'_j \right) \mathbf{V}_j^{-1}\mathbf{Z}_j\boldsymbol{\Omega}. \quad (4.18)$$

The second term in the large parentheses is due to the third term in (4.17) and will be negligible if the number m of clusters is large. The resulting simpler expression is

$$\text{Cov}(\hat{\underline{\boldsymbol{\delta}}}_j) \approx \boldsymbol{\Omega}\mathbf{Z}'_j\mathbf{V}_j^{-1}\mathbf{Z}_j\boldsymbol{\Omega}. \quad (4.19)$$

Another relevant covariance matrix contains the variances and covariances of the prediction errors. The same approximation leading to (4.19) yields

$$\text{Cov} \left(\hat{\underline{\delta}}_j - \underline{\delta}_j \right) \approx \underline{\Omega} - \underline{\Omega} \mathbf{Z}'_j \mathbf{V}_j^{-1} \mathbf{Z}_j \underline{\Omega}. \quad (4.20)$$

The variances in (4.18) and (4.19) are relevant for diagnosing properties of the residuals $\underline{\delta}_j$ and are called by Goldstein [1995] *diagnostic variances*. The variances in (4.20) are relevant for comparing residuals $\underline{\delta}_j$ and are called by Goldstein [1995] *comparative* (or *conditional*) *variances*.

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

$$\text{Cov} (\underline{\delta}_j) = \text{Cov} (\underline{\delta}_j - \hat{\underline{\delta}}_j) + \text{Cov} (\hat{\underline{\delta}}_j),$$

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 \mathbf{Z} , can be diagnosed for linearity by plots of the raw residuals $\hat{\underline{\delta}}_j$ against the level-two explanatory variables. The normality and homoscedasticity assumptions for $\underline{\delta}_j$ 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 (4.18) or (4.19), and by plotting these standardized residuals against the level-two variables. Examples are given in Goldstein [1995], Snijders and Bosker [1999], and Lewis and Langford [2001].

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

$$\hat{\underline{\delta}}_j' \left\{ \widehat{\text{Cov}} (\hat{\underline{\delta}}_j) \right\}^{-1} \hat{\underline{\delta}}_j. \quad (4.21)$$

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

With some calculations, using formula (4.7) and using the approximate covariance matrix (4.19), the standardized value (4.21) is seen to be given by

$$\hat{\underline{\delta}}_j' \left\{ \widehat{\text{Cov}} (\hat{\underline{\delta}}_j) \right\}^{-1} \hat{\underline{\delta}}_j \approx \hat{\underline{\delta}}_j^{(\text{OLS})'} \left(\hat{\sigma}^2 (\mathbf{Z}'_j \mathbf{Z}_j)^{-1} + \hat{\underline{\Omega}} \right)^{-1} \hat{\underline{\delta}}_j^{(\text{OLS})} \quad (4.22)$$

where

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

is the OLS estimate of $\underline{\delta}_j$, estimated from the residuals $\underline{\mathbf{y}}_j - \mathbf{X}_j \hat{\underline{\beta}}_j$. 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 $\hat{\sigma}^2(\mathbf{Z}'_j\mathbf{Z}_j)^{-1} + \hat{\mathbf{\Omega}}$ 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 (4.22) 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 on s d.f., as a check for outliers and for the multivariate normality of the level-two random effects.

3.4 Multivariate residuals

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

$$\underline{\mathbf{y}}_j - \mathbf{X}_j\hat{\underline{\boldsymbol{\beta}}}. \quad (4.23)$$

The covariance matrix of this residual, if we neglect the use of the estimated parameter $\hat{\underline{\boldsymbol{\beta}}}$ instead of the unknown true $\underline{\boldsymbol{\beta}}$, is given by \mathbf{V}_j in (4.15). Accordingly, the *standardized multivariate residual* is defined by

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

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

If all variables with fixed effects also have random effects, then $\mathbf{X}_j = \mathbf{Z}_j = \tilde{\mathbf{X}}_j$ as defined in Section 3.1, and $r_j = r = s$. Using (4.7), it can be proved that in this case

$$\underline{M}_j^2 = (n_j - r) \frac{s_j^2}{\sigma^2} + \hat{\underline{\boldsymbol{\delta}}}'_j \left\{ \widehat{\mathbf{Cov}}(\hat{\underline{\boldsymbol{\delta}}}_j) \right\}^{-1} \hat{\underline{\boldsymbol{\delta}}}_j, \quad (4.24)$$

in words, the standardized multivariate residual (with n_j d.f.) is the sum of the 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 (4.24) is a test statistic for the null hypothesis that the variables in \mathbf{Z}_j indeed have the effect expressed by the overall parameter estimate $\hat{\underline{\boldsymbol{\beta}}}$, i.e., the hypothesis that the variables in \mathbf{Z} and not in \mathbf{X} have only fixed (and not random) effects. This then approximately is a chi-squared variate on $r_j - r$ 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.

3.5 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 β and V_j calculated on the basis of the data set from which cluster j has been omitted. Such measures are called externally studentized residuals [Cook and Weisberg, 1982] or deletion residuals [Atkinson, 1985]. This means using the fixed parameter estimate $\hat{\beta}_{(-j)}$ obtained by estimating β from the data set from which cluster j has been omitted; and estimating (4.15) by

$$\hat{V}_{(-j)} = \mathbf{Z}_j \hat{\Omega}_{(-j)} \mathbf{Z}_j' + \hat{\Sigma}_{(-j)}, \quad (4.25)$$

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

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

$$\underline{M}_{(-j)}^2 = \left(\underline{\mathbf{y}}_j - \mathbf{X}_j \hat{\beta}_{(-j)} \right)' \hat{V}_{(-j)}^{-1} \left(\underline{\mathbf{y}}_j - \mathbf{X}_j \hat{\beta}_{(-j)} \right). \quad (4.26)$$

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

$$\hat{\underline{\delta}}_{(-j)}^{(\text{OLS})'} \left(\hat{\sigma}_{(-j)}^2 (\mathbf{Z}_j' \mathbf{Z}_j)^{-1} + \hat{\Omega}_{(-j)} \right)^{-1} \hat{\underline{\delta}}_{(-j)}^{(\text{OLS})} \quad (4.27)$$

where

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

and $\hat{\sigma}_{(-j)}^2$ 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. 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 this point of view. Two alternatives to full computation have been proposed in the literature: Lesaffre and Verbeke [1998] proposed influence statistics using an analytic approximation based on second-order Taylor expansions, and Snijders and Bosker [1999] 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.

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 [1982] and Atkinson [1985]. 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 β . Recall that $\hat{\beta}$ is the estimate obtained from the full data set, and $\hat{\beta}_{(-j)}$ 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 [1998] and Snijders and Bosker [1999] it was proposed to use the estimated covariance matrix of the estimates 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 estimate 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 [1999], but does not fit well in the analytic approach of Lesaffre and Verbeke [1998].

Denote by $\hat{\Sigma}_{F(-j)}$ the estimated covariance matrix of $\hat{\beta}_{(-j)}$ 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

$$\underline{C}_j^F = \frac{1}{r} \left(\hat{\beta} - \hat{\beta}_{(-j)} \right)' \hat{\Sigma}_{F(-j)}^{-1} \left(\hat{\beta} - \hat{\beta}_{(-j)} \right). \quad (4.28)$$

This formula has the same shape as Cook's distance.

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

$$\underline{C}_j^R = \frac{1}{p} \left(\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}_{(-j)} \right)' \hat{\boldsymbol{S}}_{R(-j)}^{-1} \left(\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}_{(-j)} \right), \quad (4.29)$$

where the analogous definitions are used for $\hat{\boldsymbol{\eta}}_{(-j)}$ and $\hat{\boldsymbol{S}}_{R(-j)}$, and p is the total number of parameters in $\boldsymbol{\eta}$. Since the parameters of the fixed and random parts are asymptotically uncorrelated [Longford, 1987], these two influence measures can be combined in the overall influence measure

$$\underline{C}_j = \frac{1}{r+p} \left(r \underline{C}_j^F + p \underline{C}_j^R \right). \quad (4.30)$$

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 (4.26). A poorly fitting group with small leverage will not do much damage to the results of the data analysis. If the model fits well, 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 (4.28-4.30) 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 (4.26) obtained from the chi-squared distribution on n_j degrees of freedom, which give information on the fit of the clusters independently of their leverage.

5. Non-linear transformations in the fixed part

In the preceding sections, various types of residuals were used to investigate possible non-linear effects of explanatory variables. The guidance provided by these methods was of an informal nature, as the non-linearity was not included in the model on which the residuals were based. The remainder of this chapter presents methods that can be used to estimate and test non-linear fixed effects of explanatory variables in a more formalized way.

We consider multilevel models for analyzing the effect of a predictor X on a response variable y under the assumption that this effect is a non-linear function $f(x)$ with an unknown functional form. The latter situation is common in longitudinal studies, for example, since the effect of time x on y is usually

complex and not well understood. Then it seems sensible to approximate $f(x)$ by a flexible function that does not require prior knowledge about $f(x)$ but still provides insight into the dependence between y and x .

In the following sections, we will successively discuss multilevel models in which a 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-1 units as repeated measurements of the subjects.

We assume that y_{ij} is the i -th response of subject j ($i = 1, \dots, n_j$; $j = 1, \dots, J$) and is generated according to the model

$$\underline{y}_{ij} = f(x_{ij}) + \sum_{k=1}^r w_{ijk} \beta_k + \sum_{k=1}^s z_{ijk} \underline{\delta}_{jk} + \underline{\epsilon}_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, J, \quad (4.31)$$

where x_{ij} , w_{ijk} , and z_{ijk} are predictors, β_k are fixed effects, $\underline{\delta}_{jk}$ are normally distributed random effects, and $\underline{\epsilon}_{ij}$ are independent normal measurement errors. The function $f(x)$ is defined on all possible values of x , not only on the observed values x_{ij} . Finally, we assume that there are T distinct values of x_{ij} , denoted by x_1, \dots, x_T in increasing order. The difference with respect to model (4.1) is that the fixed part is split into, first, a real-valued variable x with a non-linear effect, and second, R variables w_k with linear effects.

6. Polynomial model

The polynomial model for multilevel data was proposed by many authors including Goldstein [1986], Bryk and Raudenbush [1987], and Snijders [1996]. 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

$$f_{\text{pol}}(x) = \alpha_0 + \alpha_1 x + \dots + \alpha_Q x^Q.$$

The smoothness of $f_{\text{pol}}(x)$ is controlled by the degree Q . This is a linear family of functions, and therefore this model remains within the confines of the Hierarchical Linear Model, and can be estimated straightforwardly like any other such model. Another strong 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 average rate of change in $x = 0$ which sometimes is a useful parameter of a growth process. Besides, the number of parameters does not depend on the size of the data set, so that the polynomial model is easy to estimate even when the

number of distinct points T is very large. However, estimation problems may arise if x is badly scaled. Then it is advisable to rescale x by centering x around the average of observed x -values or by orthogonalizing the design matrix of the polynomial functions.

The function $f(x)$ is not always well approximated by a low-degree polynomial, however. In human growth studies, for instance, polynomials may fail to produce a smooth and accurate fit when modelling adolescent growth spurts or growth during the first year of age [Berkey and Kent, 1983]. Besides, a polynomial may exhibit non-local behavior. This means that changing one of the regression coefficients α_q leads to changes in the estimated $f_{\text{pol}}(x)$ for all values of x . A consequence 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 large value of Q fits accurately in intervals of x with many observations but may fit poorly at other values of x .

7. Regression spline model

A regression spline or piecewise polynomial [Wold, 1974] is defined as follows. A number of knots, denoted by a_1, \dots, a_M , are defined on the interval where x is observed. At each knot, two Q -th degree polynomials are connected such that the $Q - 1$ 'th derivative of the resulting function is continuous. A cubic regression spline, for instance, consists of pieces of cubic polynomial functions, its second derivative being constrained to be continuous at the knots. Regression splines are more flexible than polynomials and may produce a more accurate fit with less parameters. The reason is that because of the knots, the regression spline is better able to capture local changes in the value of $f(x)$. Regression splines are also, however, more difficult to specify than polynomials. Beside choosing the degree Q , fitting a regression spline involves choosing the number of knots M and the positions of the knots. For selecting M , an ad hoc approach can be adopted in which M is increased until an accurate fit is obtained. A more formal approach is to maximize a model summary like Akaike's Information Criterion (AIC) or the cross-validated log-likelihood [Rice and Wu, 2001]. The knots are often positioned at equally spaced points, or at quantile points of the distribution of x .

A Q 'th degree regression spline can be constructed by extending a Q 'th degree polynomial with $(M - 1)$ truncated polynomial terms $(x - c)_+^Q$, these terms being equal to $(x - c)^Q$ if $x > c$ and zero otherwise. This function $f_{\text{reg}}(x)$ can be written as

$$f_{\text{reg}}(x) = \sum_{q=0}^Q \alpha_q x^q + \sum_{m=1}^M \alpha_{Q+m} (x - a_m)_+^Q.$$

This representation is simple to understand and the α_q 's have a clear interpretation. It shows that, given that the knots are fixed, this also is a linear family of functions and therefore easy to handle, as it remains within the HLM family. However, for numerical reasons, the use of truncated polynomials is not recommendable in practical modelling, especially not when the knots lie close together. It often is better to work with a different set of basis functions. If $f_{\text{reg}}(x)$ contains one knot only, we may replace the term x^q by the term $(x - a_1)_-^q$ which equals $(x - a_1)^q$ if $x < a_1$ and 0 otherwise [Snijders and Bosker, 1999], p. 189. Note that data columns of values of $(x - a_1)_-^q$ and $(x - a_1)_+^q$ are orthogonal. If $f_{\text{reg}}(x)$ is a cubic regression spline, it is recommendable to write $f_{\text{reg}}(x)$ as a linear combination of B -splines, which are themselves piecewise cubic splines that take nonzero values over an interval with at most five knots [de Boor, 1978].

The regression spline is more flexible than the polynomial and tends to exhibit less non-local behavior. The nodes, however, are non-linear parameters and good node placement may require some trial and error. Further, if only a small number of knots are 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, one can include a large number of knots and at the same time penalize the regression coefficients such that a smooth fit is obtained [Eilers and Marx, 1996]. A limiting case is a function in which a knot is placed at each observed distinct value of x . By using as many knots as distinct data points, non-local behavior is avoided. Splines of this type will be discussed in the next section.

8. Smoothing spline model

The cubic smoothing spline, denoted by $f_{\text{css}}(x)$, is a cubic regression spline with a knot at the ordered observed distinct values x_1, \dots, x_T and constrained to be linear for values beyond the boundary knots x_1 and x_T . The degree of smoothness is regulated by a roughness penalty defined as

$$-\frac{1}{2}\lambda \int_{x_0}^{x_{T+1}} \{f''(x)\}^2 dx, \quad (4.32)$$

where λ is a nonnegative smoothing parameter determining the degree of smoothing, $x_0 \leq x_1$, and $x_{T+1} \geq x_T$. This penalty reflects roughness in an intuitively appealing way: smooth functions tend to have low absolute curvature meaning that the value of the second derivative $f''(x)$ is close to zero over the interval of observed x .

The following basic properties of smoothing splines can be found in the literature on this topic, such as Green and Silverman [1994]. The fitted cubic smoothing spline is obtained by maximizing the penalized log-likelihood, i.e., the sum of the log-likelihood and the roughness penalty. An additional

constraint to ensure that $f(x)$ is a cubic smoothing spline does not have to be included because among all functions $f(x)$ with continuous second derivatives, the unique minimizer of the penalized log-likelihood is a cubic smoothing spline. If we substitute the cubic smoothing spline $f_{\text{css}}(x)$ in the roughness penalty, we can simplify (4.32) to

$$-\frac{1}{2}\lambda \mathbf{f}'_{\text{css}} \mathbf{K} \mathbf{f}_{\text{css}},$$

where \mathbf{f}_{css} is the vector of values of $f_{\text{css}}(x)$ at the distinct points x_1, \dots, x_T . The $T \times T$ matrix \mathbf{K} equals

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

where \mathbf{Q} is a $T \times (T - 2)$ matrix having entries $q_{i+1,i} = 1/(x_{i+1} - x_i)$, $q_{i-1,i} = 1/(x_i - x_{i-1})$, $q_{i,i} = -(q_{i-1,i} + q_{i+1,i})$ for $i = 2, \dots, T - 1$, and zero otherwise. The $(T - 2) \times (T - 2)$ matrix \mathbf{R} is symmetric tridiagonal with diagonal entries $r_{i,i} = \frac{1}{3}(x_{i+1} - x_i) + \frac{1}{3}(x_{i+2} - x_{i+1})$ for $i = 1, \dots, T - 2$. 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 $i = 1, \dots, T - 3$.

8.1 Estimation

We discuss the Expectation Maximization (EM) algorithm [Dempster et al., 1977] and the Restricted Iterative Generalized Least Squares (RIGLS) algorithm [Goldstein, 1989]. The model parameters to be estimated are the vector of spline values \mathbf{f}_{css} , the vector of other fixed regression coefficients $\boldsymbol{\beta}$, the level-one variance σ^2 , and the level-two variance parameters $\boldsymbol{\xi}$.

EM algorithm. The EM algorithm consists of an E-step and an M-step. In the E-step, we define the complete-data log likelihood, i.e. the log joint density function of responses \mathbf{y} and the vector of random coefficients $\boldsymbol{\delta}$ (treated in this algorithm as missing data), and maximize it conditional on \mathbf{y} . The complete-data log likelihood is given by

$$\begin{aligned} & \log p(\mathbf{y}, \boldsymbol{\delta} | \mathbf{f}_{\text{css}}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\xi}) \\ &= \log p(\mathbf{y} | \mathbf{f}_{\text{css}}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) + \log p(\boldsymbol{\delta} | \boldsymbol{\xi}) \\ &= -\frac{1}{2} \sum_j \left\{ n_j \log(\sigma^2) \right. \\ & \quad \left. - (\mathbf{y}_j - \mathbf{N}_j \mathbf{f}_{\text{css}} - \mathbf{W}_j \boldsymbol{\beta} - \mathbf{Z}_j \boldsymbol{\delta}_j)' (\mathbf{y}_j - \mathbf{N}_j \mathbf{f}_{\text{css}} - \mathbf{W}_j \boldsymbol{\beta} - \mathbf{Z}_j \boldsymbol{\delta}_j) / \sigma^2 \right. \\ & \quad \left. + \log \det(\boldsymbol{\Omega}(\boldsymbol{\xi})) - \boldsymbol{\delta}_j' (\boldsymbol{\Omega}(\boldsymbol{\xi}))^{-1} \boldsymbol{\delta}_j \right\}, \end{aligned}$$

with \mathbf{N}_j an $n_j \times T$ matrix of zeros and ones. Each row of \mathbf{N}_j contains a single one at the entry t for which $x_{ij} = x_t$.

In the M-step, we maximize the sum of the expected complete-data log likelihood and the roughness penalty with respect to the model parameters \mathbf{f}_{css} , $\boldsymbol{\beta}$, σ^2 , and $\boldsymbol{\xi}$. The M-step is computationally expensive if the number of distinct time points T is large. Then it is better to update the estimates of \mathbf{f}_{css} and $\boldsymbol{\beta}$ sequentially. If maximize the penalized function with respect to \mathbf{f}_{css} only, we obtain the updated estimate

$$\hat{\mathbf{f}}_{\text{css}} = \left(\sum_j (\mathbf{N}'_j \mathbf{N}_j) + \hat{\sigma}^2 \lambda \mathbf{K} \right)^{-1} \sum_j (\mathbf{y}_j - \mathbf{W}_j \hat{\boldsymbol{\beta}} - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j),$$

where $\hat{\boldsymbol{\delta}}_j$ are EB level-two residuals and the current estimates are filled in for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Analogously, if we maximize with respect to $\boldsymbol{\beta}$ only, we obtain

$$\hat{\boldsymbol{\beta}} = \left(\sum_j \mathbf{W}'_j \mathbf{W}_j \right)^{-1} \sum_j (\mathbf{y}_j - \mathbf{N}_j \hat{\mathbf{f}}_{\text{css}} - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j).$$

The two steps are computationally cheap: to update the estimates of \mathbf{f}_{css} and $\boldsymbol{\beta}$, $c_1 T$ numerical operations are required, whereas joint maximization of \mathbf{f}_{css} and $\boldsymbol{\beta}$ requires $c_2 T^3$ numerical operations (c_1 and c_2 are constants). This algorithm where the M-step is replaced by two sequential steps is known as the EC(onditional)M algorithm [Meng and Rubin, 1993]. The two sequential steps can be regarded as backfitting steps as described by Hastie and Tibshirani [1990], p. 91.

RIGLS estimation. We reformulate model (4.31) as a two-level model with crossed random effects and estimate the parameters by the restricted maximum likelihood (REML) method (Speed [1991]; Wang [1998]; Zhang et al. [1998]). To derive the crossed effects model, we write $f(x_{ij})$ in (4.31) as

$$f(x_{ij}) = \gamma_0 + \gamma_1 x_{ij} + N_{ij} \mathbf{Q} (\mathbf{Q}' \mathbf{Q})^{-1} \mathbf{L} \boldsymbol{\delta}, \quad (4.33)$$

where γ_0 and γ_1 are scalars, $\boldsymbol{\delta}$ is an $(T - 2) \times 1$ vector, \mathbf{L} satisfies $\mathbf{L} \mathbf{L}' = \mathbf{R}$, and N_{ij} is an $1 \times T$ incidence row vector of $T - 1$ zeros and a single value 1 at the entry t for which x_{ij} equals x_t [Green, 1987]. We substitute (4.33) in model (4.31) and assume that γ_0 and γ_1 are fixed effects and $\boldsymbol{\delta}$ is a normal crossed random effect with mean zero and covariance matrix $(1/\lambda) \mathbf{I}_T$.

This model can be estimated by means of the RIGLS algorithm implemented in software packages like *MLwiN* [Goldstein et al., 1998] and *SAS* [Littell et al., 1996]. An estimate of \mathbf{f}_{css} is obtained by substituting the (restricted) maximum likelihood estimates of γ_0 and γ_1 and the empirical Bayes estimate of $\boldsymbol{\delta}$ in (4.33). This estimator is the best linear unbiased predictor (BLUP) as noted by Speed [1991] and also is the maximizer of the penalized log-likelihood.

The formulation of the crossed effects model is attractive because it allows us to estimate \mathbf{f}_{css} using existing software. However, since the number of crossed

random effects is equal to $T - 2$, the estimation is numerically demanding if there are many distinct points. In that case, the ECM algorithm is preferable. We further note that the crossed effects model is useful for computational purposes, but is not a realistic description of how the data are generated. For instance, the presence of crossed random effects suggests that measurements from different subjects are correlated, but this often is not a plausible assumption.

8.2 Inferences

A common approach to drawing inferences is to construct a pointwise correct confidence interval around the estimator $\hat{\underline{f}}_{\text{css}}$. This requires an estimate for the variance of $\hat{\underline{f}}_{\text{css}}(x)$ for each x , which can be calculated from the crossed effects model. If the crossed effects model is postulated, the covariance matrix of $\hat{\underline{f}}_{\text{css}}$ equals

$$\text{Cov}(\hat{\underline{f}}_{\text{css}}) = \text{Cov}\left(\hat{\underline{\gamma}}_0 \boldsymbol{\nu}_T + \hat{\underline{\gamma}}_1 \mathbf{x} + \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{L}\hat{\underline{\delta}}\right), \quad (4.34)$$

where $\boldsymbol{\nu}_T$ is a $T \times 1$ vector of ones. We see from (4.34) that we have an estimator of $\text{Cov}(\hat{\underline{f}}_{\text{css}})$ once we have an estimate for the covariance matrix of $\hat{\underline{\gamma}}_0$, $\hat{\underline{\gamma}}_1$, and $\hat{\underline{\delta}}$. Several estimators exist for the variance of $\hat{\underline{\delta}}$. If we use the comparative variance (4.20), the estimator of $\text{Cov}(\hat{\underline{f}}_{\text{css}})$ has a nice Bayesian interpretation: it is equal to the posterior variance of $\underline{f}_{\text{css}}$ under a Bayesian model with a Gaussian process prior for \underline{f} that is improper with respect to intercept and slope (Wang [1998]; Zhang et al. [1998]). This model was put forward by Wahba [1983] and serves as a Bayesian justification for the use of roughness penalties. In some software packages such as *MLwiN*, the covariance between $\hat{\underline{\delta}}$ and $(\hat{\underline{\gamma}}_0, \hat{\underline{\gamma}}_1)$ cannot be easily estimated. However, the design matrices for (γ_0, γ_1) and $\underline{\delta}$ are orthogonal if the points at which the measurements are taken are common to all subjects. Therefore, the precision of the estimator of $\text{Cov}(\hat{\underline{f}})$ is in general not substantially affected by the omission of the covariance between $(\hat{\underline{\gamma}}_0, \hat{\underline{\gamma}}_1)$ and $\hat{\underline{\delta}}$.

The Bayesian covariance matrix of $\hat{\underline{f}}_{\text{css}}$ can also be written in the following simple form

$$\text{Cov}_B(\hat{\underline{f}}_{\text{css}}) = (\mathbf{U} + \lambda\mathbf{K})^{-1}, \quad (4.35)$$

where

$$\mathbf{U} = \sum_j \mathbf{N}'_j \mathbf{V}_j^{-1} \mathbf{N}_j - \sum_j \mathbf{N}'_j \mathbf{V}_j^{-1} \mathbf{X}_j \left(\sum_j \mathbf{X}'_j \mathbf{V}_j^{-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}'_j \mathbf{V}_j^{-1} \mathbf{N}_j,$$

with \mathbf{V}_j given in (4.6) and \mathbf{N}_j defined as the matrix with rows N_{ij} . Besides the Bayesian covariance matrix, a frequentist covariance matrix can be obtained as

well if we assume that f_{css} is an unknown fixed function and $\text{Cov}(\underline{\mathbf{y}}_j) = \mathbf{V}_j$. The covariance matrix then is given by

$$\text{Cov}_F(\hat{\underline{\mathbf{f}}}_{\text{css}}) = (\mathbf{U} + \lambda\mathbf{K})^{-1} \mathbf{U} (\mathbf{U} + \lambda\mathbf{K})^{-1}. \quad (4.36)$$

Zhang et al. [1998] and Lin and Zhang [1999] compare the two estimators 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 performs slightly better.

Inferences can also be made by performing a likelihood ratio or F -test to examine, for instance, whether a model with smoothing spline $f(x)$ fits better than a model with a linear effect for x . Hastie and Tibshirani [1993] (p. 65) provide some approximate F -tests based on residual sums of squares and Cantoni and Hastie [2002] present an exact but more expensive test.

A third test against a linear effect of X is the score test for $H_0 : \gamma_1 = 0$ for model (4.33) under the additional assumption that the effect is indeed linear ($\delta = 0$). This test was considered in Guo [2002]. The score test is computationally cheap because REML estimates of the alternative model are not required. The test is based on the one-step REML 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 has good power properties also in a small sample setting [Berkhof and Snijders, 2002b]. For testing against an unspecified but monotone effect of X , this test against a linear effect may be expected to have good power also against most non-linear effects.

8.3 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 expected predictive log-likelihood which is the expected log-likelihood of a new vector of observations $\underline{\mathbf{y}}^*$ at the penalized likelihood estimators of the model parameters $\hat{\underline{\mathbf{f}}}_{\text{css}}$, $\hat{\underline{\boldsymbol{\beta}}}$, σ^2 , and $\hat{\underline{\boldsymbol{\xi}}}$. 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 [Rice and Silverman, 1991].

A drawback of cross-validation is that it is computationally expensive when the data set contains many subjects. One can also estimate the expected predictive log-likelihood by the sum of the log-likelihood and the trace of the matrix \mathbf{A} that maps $\underline{\mathbf{y}}$ on the estimator $\hat{\underline{\mathbf{f}}}_{\text{css}} = \mathbf{A}\underline{\mathbf{y}}$ (Hastie and Tibshirani [1990] p. 52, Green and Silverman [1994] p. 37). This estimator, named Mallows' C_p , is cheap and unbiased if the (co)variance parameters σ^2 and $\underline{\boldsymbol{\xi}}$ are known. For uncorrelated data, the unbiasedness proof is provided by Hastie and Tibshirani

[1990], p. 48. The proof in the case of multilevel data is analogous. In practice, σ^2 and ξ are unknown and have to be replaced by estimators $\hat{\sigma}^2$ and $\hat{\xi}$. A possible choice for $\hat{\sigma}^2$ and $\hat{\xi}$ is the REML estimator in a model without smoothing ($\lambda = 0$). This estimator is cheap and consistent also when the true model is defined by a smooth function $f(x)$.

A drawback of the above methods is that λ is not treated as a model parameter but as an external variable. This complicates the inference about the (co)variance parameters ξ . The smoothing parameter becomes a model parameter if we adopt the Bayesian model of Wahba [1983]. We can now estimate λ by maximizing the likelihood with respect to σ^2 and ξ and λ [Wahba, 1985], Zhang et al. [1998] after having integrated out the other model parameters. An attractive feature of this so-called generalized maximum likelihood (GML) method is that the maximizers $\hat{\sigma}^2$, $\hat{\xi}$ and $\hat{\lambda}$ are the REML estimators of the crossed effects model. Therefore, if we estimate the parameters of the crossed effects model by means of RIGLS implemented in *MLwiN* [Goldstein et al., 1998], we obtain penalized maximum likelihood estimates for the model parameters and a GML estimate for the smoothing parameter.

9. 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* [Goldstein et al., 1998] and *Gauss* [Systems, 1994]. The data set used is described in Snijders and Bosker [1999]. It contains language test scores of 2287 pupils within 131 elementary schools. We modeled the test score (Y) as a function of the centered IQ of the pupil (IQ), the gender of the pupil (SEX), the school average of IQ ($\overline{\text{IQ}}$), and the social economic status (SES) of the pupil. A non-linear effect was assumed for IQ and linear effects were assumed for the other predictors. Note that in most applications of models with functional non-linear effects, some time variable is the ordering principle but an ordering according to any other unidimensional variable is possible as well. We further accounted for between-school differences by including a random intercept and a random slope of IQ at level two. Finally, the level one measurement errors are assumed to be homoskedastic and uncorrelated. The model can be written as

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

We present the fit of three different functions $f(\text{IQ})$ in Figure 4.3. The first function is a cubic polynomial. (We also considered a fourth-degree polynomial but this did not yield a further improvement in fit.) The second function is a quadratic regression spline with a knot at zero. This function was considered

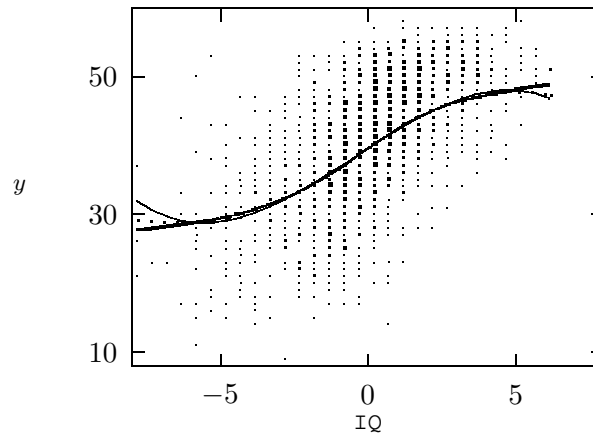


Figure 4.3. Language test score (y) against centered IQ score: raw data, cubic polynomial estimate (thin), quadratic regression spline estimate (dashed), and smoothing spline estimate (bold).

by Snijders and Bosker [1999], p. 113 as a flexible and parsimonious alternative for the polynomial function. The third function is a cubic smoothing spline, the smoothness of which is determined by applying the GML method. (Cross-validation, Mallows C_p , and GML rendered similar values for the smoothing parameter: $\lambda_{\text{GML}} = 1.6$, $\lambda_{C_p} = 1.6$, $\lambda_{\text{CV}} = 2.0$.) We see that the three fitted functions tell the same story: 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 are decreasing at low and high values of IQ.

We also estimated the pointwise standard errors of the fitted functions. These are presented in Figure 4.4. We see that the standard errors of the fitted functions are very similar. Data are sparse at the left and right ends of the window (Figure 4.4) 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 which is in accordance with (4.35) and (4.36) [Zhang et al., 1998].

10. Generalizations

We only considered a two-level model for normal responses. The model can be generalized 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. We can also specify a model with two functional effects, $f(x)$ and $g(v)$, where x and

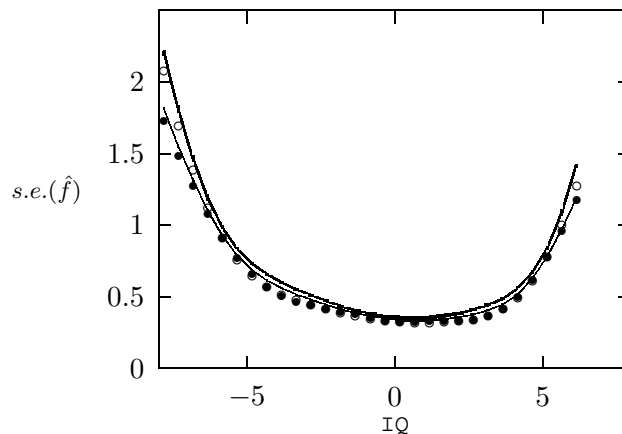


Figure 4.4. Standard errors of the cubic polynomial estimate (open circle) and the quadratic regression spline estimate (closed circle), and Bayesian (bold line) and frequentist (thin line) standard errors of the smoothing spline estimate.

v are underlying variables. This model is named an additive model and was put forward by Hastie and Tibshirani [1990]. Algorithms for estimating additive multilevel models with cubic smoothing splines are provided by Lin and Zhang [1999]. A related but slightly different model is a model with functional effects $f(x)$ and $g(x)h(x)$, where $g(x)$ is the functional effect of predictor $h(x)$. This model is known as a varying coefficient model [Hastie and Tibshirani, 1993]. A multilevel extension of the model is presented by Hoover et al. [1998]. The additive and varying coefficient models can be formulated as crossed effects models with a separate crossed 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 (Fan and Zhang [2000], Chiang et al. [2001]).

So far, we only discussed functional effects in the fixed part of the model but the random part of the model may contain functional effects as well. Multilevel models with polynomial random effects are discussed by Goldstein [1986], Bryk and Raudenbush [1987], and Snijders [1996]. Rice and Wu [2001] present a multilevel model where the design matrix of the random effects is a set of B-spline basis functions. James et al. [2000] present a similar model where the design matrix of the random effects consists of B-splines rather than B-spline basis functions. Note that B-splines are linear combinations of B-spline basis functions. Rice and Silverman [1991] propose a model where the design matrix consists of cubic smoothing splines. Their model is applicable only when the points at which measurements are taken are common to all level two units. Berkhof and Snijders [2002a] propose an extension of the model of Rice and

Silverman [1991] that allows for missing data. Finally, Brumback and Rice [1998] and Guo [2002] present models where the design matrix consists of natural cubic spline basis functions.

The above models with splines in the random part can be classified according to the dimensionality of the covariance matrix. The models of Brumback and Rice [1998], Rice and Wu [2001], and Guo [2002] have a high-dimensional level two covariance matrix whereas the dimension of the covariance matrix is usually not larger than five for the models of Rice and Silverman [1991], James et al. [2000], and Berkhof and Snijders [2002a]. The latter models can be regarded as principal component models where the splines in the random part describe the main sources of variation among the individual curves.

References

- A.C. Atkinson. *Plots, Transformations, and Regression*. Clarendon Press, Oxford, 1985.
- B.H. Baltagi. *Econometric Analysis of Panel Data*. Wiley, Chichester, 1995.
- M.S. Bartlett and D.G. Kendall. The statistical analysis of variances-heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society*, Supplement 8:128–138, 1946.
- C.S. Berkey and R.L. Jr. Kent. Longitudinal principal components and non-linear regression models of early childhood growth. *Annals of Human Biology*, 10:423–536, 1983.
- J. Berkhof and T.A.B. Snijders. Reduced rank smoothing spline models for unbalanced curve data. Submitted, 2002a.
- J. Berkhof and T.A.B. Snijders. Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, 26:133–152, 2002b.
- B.A. Brumback and J.A. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, 93:961–994, 1998.
- A.S. Bryk and S.W. Raudenbush. Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101:147–158, 1987.
- A.S. Bryk and S.W. Raudenbush. *Hierarchical Linear Models, Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA, 1992.
- E. Cantoni and T. Hastie. Degrees-of-freedom tests for smoothing splines. *Biometrika*, 89:251–263, 2002.

- C.T. Chiang, J. Rice, and C. Wu. Smoothing splines estimation for varying coefficient models with repeatedly dependent variable. *Journal of the American Statistical Association*, 96:605–619, 2001.
- R.D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman & Hall, New York and London, 1982.
- C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.
- P.H.C. Eilers and B.D. Marx. Flexible smoothing with splines and penalties (with discussion). *Statistical Science*, 11:89–121, 1996.
- 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.
- H. Goldstein. Efficient statistical modelling of longitudinal data. *Annals of Human Biology*, 13:129–141, 1986.
- H. Goldstein. Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76:622–623, 1989.
- H. Goldstein. *Multilevel Statistical Models*. Edward Arnold, London, second edition, 1995.
- 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, 1998.
- P.J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55:245–259, 1987.
- P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.
- W. Guo. Functional mixed effects model. *Biometrics*, 58:121–128, 2002.
- T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.
- T. Hastie and R. Tibshirani. Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B*, 55:757–796, 1993.

- J.A. Hausman. Specification tests in econometrics. *Econometrica*, 46:1251–1271, 1978.
- J.A. Hilden-Minton. *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*. PhD thesis, Department of Mathematics, University of California, Los Angeles, 1995.
- 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.
- 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.
- G. James, T. Hastie, and C.A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000.
- I.H. Langford and T. Lewis. Outliers in multilevel data. *Journal of the Royal Statistical Society, Ser. A*, 161:121–160, 1998.
- E. Lesaffre and G. Verbeke. Local influence in linear mixed models. *Biometrics*, 54:570–582, 1998.
- T. Lewis and I.H. Langford. *Outliers, robustness and the detection of discrepant data*, volume Multilevel Modelling of Health Statistics, chapter 6, pages 75–91. Wiley, New York, 2001.
- 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.
- R.C. Littell, G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. *SAS System for Mixed Models*. SAS Institute, Cary, NC, 1996.
- N.T. Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74:812–827, 1987.
- N.T. Longford. *Random Coefficient Models*. Oxford University Press, New York, 1993.
- X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, 80:267–278, 1993.
- C.R. Rao. *Linear Statistical Inference*. Wiley, New York, second edition, 1973.

- J. Rice and C. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57:253–259, 2001.
- 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.
- M.H. Seltzer, W.H.M. Wong, and A.S. Bryk. Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational Statistics*, 21:131–167, 1996.
- T.A.B. Snijders. Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, 30:405–426, 1996.
- T.A.B. Snijders and R.J. Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications, London, 1999.
- 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.
- Aptech Systems. *Gauss*. Aptech Systems, Maple Valley, Washington, 1994.
- 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.
- G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer, New York, 2000.
- G. Wahba. Bayesian “confidence” intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, 45:133–150, 1983.
- 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.
- Y. Wang. Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, 60:159–174, 1998.
- 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.
- S. Wold. Spline functions in data analysis. *Technometrics*, 16:1–11, 1974.
- 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.