TECHNICAL REPORT

0663

# FORMAL AND INFORMAL MODEL SELECTION WITH INCOMPLETE DATA

VERBEKE G., and G. MOLENBERGHS

IAP STATISTICS NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

# Formal and Informal Model Selection with Incomplete Data

Geert Verbeke [*] Geert Molenberghs [†]

## Summary

Model selection and assessment with incompletely data pose challenges in addition to the ones encountered with complete data. There are two main reasons for this. First, many models describe characteristics of the complete data, in spite of the fact that only an incomplete subset is observed. Direct comparison between model and data is then less than straightforward. Second, many commonly used models are more sensitive to assumptions than in the complete data situation and some of their properties vanish when they are fitted to incomplete, unbalanced data. These and other issues are brought forward using two key examples, one of a continuous and one of a categorical nature. We argue that model assessment ought to consist of two parts: (i) assessment of a model's fit to the observed data and (ii) assessing the sensitivity of inferences to unverifiable assumptions, i.e., to how a model described the unobserved data given the observed ones.

**Keywords and phrases :** Interval of ignorance; Linear mixed model; Missing at random; Missing not at random; Multivariate normal; Sensitivity analysis.

**AMS 1991 subject classifications.** Primary 62-07; secondary 62P10

**Acknowledgment of support:** We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

[*] Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium

[†] Center for Statistics, Hasselt University, Agoralaan 1, B3590 Diepenbeek, Belgium

# 1 Introduction

In many longitudinal and multivariate settings, not all measurements envisaged in the design stage are taken in actual practice. It is important to reflect on the nature and implications of such incompleteness, or missingness, and properly accommodate it in the modeling process. Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Afifi and Elashoff 1966, Hartley and Hocking 1971). Nowadays, general algorithms such as expectation-maximization (EM) (Dempster, Laird, and Rubin 1977), and data imputation and augmentation procedures (Rubin 1987), combined with powerful computing resources and flexible implementations in standard software have largely resolved the computational difficulties. There remains the difficult and important question of assessing the impact of missing data on subsequent statistical inference.

When referring to the missing-value, or non-response, process we will use terminology of Little and Rubin (2002, Chapter 6). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR).

Given MAR, a valid analysis that ignores the missing value mechanism can be obtained, within a likelihood or Bayesian framework, provided the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, the so-called parameter distinctness condition. This situation is termed ignorable by Rubin (1976) and Little and Rubin (2002) and leads to considerable simplification in the analysis (Diggle 1989, Verbeke and Molenberghs 2000). There is a strong trend, nowadays, to prefer this kind of analyses, in the likelihood context also termed *direct-likelihood* analysis, over *ad hoc* methods such as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation (Molenberghs *et al* 2004, Mallinckrodt *et al* 2003ab, Jansen *et al* 2006a). Practically, it means conventional tools for longitudinal and multivariate data, such as the linear and generalized linear mixed-effects models

(Verbeke and Molenberghs 2000, Molenberghs and Verbeke 2005) can be used in exactly the same way as with complete data. Such software tools as the SAS procedures MIXED, NLMIXED, and GLIMMIX facilitate this paradigm shift.

One should be aware that, in spite of the flexibility and elegance a direct-likelihood method brings, there are fundamental issues when selecting a model and assessing its fit to the observed data, which do not occur with complete data. Such issues are the central theme of this paper; already in the MAR case, but they are compounded further under MNAR.

Indeed, one can never fully rule out MNAR, in which case the missingness mechanism needs to be modeled alongside the mechanism generating the responses. In the light of this, one approach could be to estimate from the available data the parameters of a model representing a MNAR mechanism. It is typically difficult to justify the particular choice of missingness model, and the data do not necessarily contain information on the parameters of the particular model chosen (Jansen *et al* 2006b). For example, different MNAR models may fit the observed data equally well, but have quite different implications for the unobserved measurements, and hence for the conclusions to be drawn from the respective analyses. Without additional information one can only distinguish between such models using their fit to the observed data, and so goodness-of-fit tools alone do not provide a relevant means of choosing between such models. This implies the necessity of sensitivity analysis when assessing the quality of inferences from incomplete data, defined, in a broad way, as an instrument to assess the impact on statistical inferences from varying the, often untestable, assumptions in an MNAR model. Overviews may be found in Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007).

The ideas will be developed by means of two running examples, which are introduced in Section 2, along with the model families to be used. Initial analyses are presented as well. A number of issues arising when analyzing such incomplete data, under MAR as well as MNAR, are enlisted in Section 3. Ways of tackling the problems are the subject of Section 4.

# 2 Running Examples and Their Initial Analyses

In Section 2.1, the orthodontic growth data are introduced, together with the linear mixed models used for their analysis. Similarly, Section 2.2 is devoted to the introduction of the Slovenian Public Opinion Survey, along with the model family of Baker, Rosenberger, and DerSimonian (1992).

## 2.1 The Orthodontic Growth Data

These data, introduced by Pothoff and Roy (1964), contain growth measurements for 11 girls and 16 boys. For each subject, the distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14. The data were used by Jennrich and Schluchter (1986) to illustrate estimation methods for unbalanced data, where unbalancedness is now to be interpreted in the sense of an unequal number of boys and girls. Individual profiles and sex group by age means are plotted in Figure 1.

Little and Rubin (2002) deleted 9 of the $[(11+16)\times 4]$ observations, thereby producing 9 incomplete subjects with a missing measurement at age 10. Their missingness generating mechanism was such that subjects with a low value at age 8 are more likely to have a missing value at age 10. The data are presented in Table 1. The measurements that were deleted are marked with an asterisk. We first focuss on the analysis of the original complete data set.

Jennrich and Schluchter (1986), Little and Rubin (2002), and Verbeke and Molenberghs (1997, 2000) each fitted the same eight models, which can be expressed within the general linear mixed models family (Verbeke and Molenberghs 2000):

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where

$$\begin{aligned}\boldsymbol{b}_i &\sim N(\mathbf{0}, D),\\ \boldsymbol{\varepsilon}_i &\sim N(\mathbf{0}, \Sigma_i),\end{aligned}$$

and $\boldsymbol{b}_i$ and $\varepsilon_i$ are statistically independent. Here, $\boldsymbol{Y}_i$ is the $(4\times 1)$ response vector, $X_i$ is a $(4\times p)$

design matrix for the fixed effects, $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients, $Z_i$ is a $(4 \times q)$ design matrix for the random effects, $\boldsymbol{b}_i$ is a $(q \times 1)$ vector of normally distributed random parameters, with covariance matrix $D$, and $\boldsymbol{\varepsilon}_i$ is a normally distributed $(4 \times 1)$ random error vector, with covariance matrix $\Sigma$. Estimation and inference is traditionally obtained from likelihood principles based on the marginal distribution $\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i)$.

In our example, every subject contributes exactly four measurements at exactly the same time points. It is therefore possible to drop the subscript $i$ from the error covariance matrix $\Sigma_i$ unless, for example, sex is thought to influence the residual covariance structure. The random error $\boldsymbol{\varepsilon}_i$ encompasses both within-subject variability and serial correlation. The mean $X_i\boldsymbol{\beta}$ will be a function of age, sex, and/or the interaction between both.

Table 2 summarizes model fitting and comparison for the eight models originally considered by Jennrich and Schluchter (1987). The initial Model 1 assumes an unstructured group by time model, producing eight mean parameters. In addition, the variance-covariance matrix is left unstructured, yielding an additional ten parameters. First, the mean structure is simplified, followed by the covariance structure. Models 2 and 3 consider the mean profiles to be non-parallel and parallel straight lines, respectively. While the second model fits adequately, the third one does not, based on conventional likelihood ratio tests. Thus, the crossing lines will be retained. Models 4 and 5 assume the variance-covariance structure to be of a banded (Toeplitz) and first-order auto-regressive (AR(1)) type, respectively. Model 6 assumes the covariance structure to arise from correlated random intercepts and random slopes. In Model 7, a compound-symmetry structure is assumed, which can be seen as the marginalization of a random-intercepts model. Finally, Model 8 assumes uncorrelated measurements. Of these, Models 4, 6, and 7 are well-fitting. Model 7, being the most parsimonious one, will be retained.

Let us now fit the same eight models to the trimmed, incomplete, version of the dataset, as presented by Little and Rubin (2002), using direct-likelihood methods. This implies the same models are fitted with the same software tools, but now to a reduced set of data. The results are summarized in Table 2 as well. Note that the same Model 7 is selected. A quite different picture would emerge, were simple, ad hoc, methods used (Molenberghs and Kenward, 2007). Table 3 presents finally selected models based on a complete case (CC) analysis, on last observation carried forward (LOCF), as well as on unconditional and conditional mean imputation. A complete case analysis completely ignores the children with not all 4 observations measured. Under LOCF, the missing observations are imputed with the last observed measurement, i.e., the observation measured at age 8. Conditional and unconditional mean imputation replace the missing observations by the gender-specific average, or the predicted value conditionally on the observed outcomes, respectively. Note that none of these approaches recover Model 7, even though CC and unconditional mean imputation lead to a slight modification of this model. In contract, LOCF and conditional mean imputation produce a much more complicated and therefore cumbersome final model. This illustrates the point that simple methods such as CC and LOCF, as well as other simple imputation methods, can be quite distorting, whereas direct likelihood retains its validity under MAR. One might argue that the price to pay is the need to fit a model to the entire longitudinal sequence, even in circumstances where scientific interest focuses on the last planned measurement occasion. For continuous data, an obvious choice for such a full longitudinal model is the linear mixed model. However, for balanced longitudinal data, where the number of subjects is sufficiently large compared to the number of times, a full multivariate normal, such as our Model 1, can often be considered, not making assumptions beyond the ones made by, say, multivariate analysis of variance (MANOVA), ANOVA per time point or, equivalently, a $t$ test per time point. This is illustrated in Table 4, using Model 1 fitted to the complete and trimmed growth data. Means for boys at the ages 8 and 10 are displayed. Whenever the data are balanced, the means are the same regardless of which estimation method is used. Standard errors are asymptotically the same and even in a small sample like the one considered here, differences are negligible. Note that CC overestimates the means since the subjects removed from analysis have lower means than average, and LOCF underestimates the mean at age 10, since the age 8 measurement is carried forward.

When the observed data are analyzed, it is clear that the results from the direct likelihood analyses, valid under MAR, diverge from the frequentist MANOVA and ANOVA analyses, which are valid only under MCAR. MANOVA effectively reduces to CC, due to its inability to take incomplete sequences into account. ANOVA produces correct inferences only at measurement occasions with complete data.

## 2.2 The Slovenian Public Opinion Survey

In 1991 Slovenians voted for independence from former Yugoslavia in a plebiscite. To prepare for this result, the Slovenian government collected data in the Slovenian Public Opinion Survey (SPO), a month prior to the plebiscite. Rubin, Stern, and Vehovar (1995) studied the three fundamental questions added to the SPO and, in comparing it to the plebiscite's outcome, drew conclusions about the missing data process.

The three questions added were: (1) Are you in favour of Slovenian independence? (2) Are you in favour of Slovenia's secession from Yugoslavia? (3) Will you attend the plebiscite? In spite of their apparent equivalence, questions (1) and (2) are different since independence would have been possible in confederal form as well and therefore the secession question is added. Question (3) is highly relevant since the political decision was taken that not attending was treated as an effective NO to question (1). Thus, the primary estimand is the proportion $\theta$ of people that will be considered as voting YES, which is the fraction of people answering yes to both the attendance and independence question. The raw data are presented in Table 5.

The data were used by Molenberghs, Kenward, and Goetghebeur (2001) to illustrate their sensitivity analysis tool, the interval of ignorance. Molenberghs *et al* (2006) used the data to exemplify results about the relationship between MAR and MNAR models. An overview of various analyses can be found in Molenberghs and Kenward (2007). These authors used the model proposed by Baker, Rosenberger, and DerSimonian (1992) for the setting of two-way contingency tables subject to non-monotone missingness. Such data take the form of counts $Z_{r_1,r_2,jk}$, where $j,k = 0,1$ reference the two categories and $r_1, r_2 = 0,1$ are the missingness indicators for each. The corresponding probabilities are $\nu_{r_1,r_2,jk}$, describing a four-way classification, and modeled as:

$$\begin{aligned} \nu_{10,jk} &= \nu_{11,jk}\beta_{jk}, \\ \nu_{01,jk} &= \nu_{11,jk}\alpha_{jk}, \\ \nu_{00,jk} &= \nu_{11,jk}\alpha_{jk}\beta_{jk}\gamma. \end{aligned} \tag{2}$$

The $\alpha$ ($\beta$) parameters describe missingness in the independence (attendance) question, and $\gamma$ captures the interaction between both. The subscripts are missing from $\gamma$ since Baker, Rosenberger, and DerSimonian (1992) have shown that this quantity is independent of $j$ and $k$ in every identifiable model. These authors considered nine models, based on setting $\alpha_{jk}$ and $\beta_{jk}$ constant in one or more indices:

$$\begin{array}{lclllcl} \text{BRD1} & : & (\alpha,\beta) & \text{BRD4} & : & (\alpha,\beta_k) & \text{BRD7} \; : \; (\alpha_k,\beta_k) \\ \text{BRD2} & : & (\alpha,\beta_j) & \text{BRD5} & : & (\alpha_j,\beta) & \text{BRD8} \; : \; (\alpha_j,\beta_k) \\ \text{BRD3} & : & (\alpha_k,\beta) & \text{BRD6} & : & (\alpha_j,\beta_j) & \text{BRD9} \; : \; (\alpha_k,\beta_j). \end{array}$$

Interpretation is straightforward, for example, BRD1 is MCAR, and in BRD4 missingness in the first variable is constant, while missingness in the second variable depends on its value. BRD6–BRD9 saturate the observed data degrees of freedom, while the lower numbered ones do not, leaving room for a non-trivial model fit to the observed data.

Rubin, Stern, and Vehovar (1995) conducted several analyses of the data. Their main emphasis was in determining the proportion $\theta$ of the population that would attend the plebiscite and vote for independence. The three other combinations of these two binary outcomes would be treated as voting "no". Their estimates are reproduced in Table 6.

The pessimistic (optimistic) bounds are obtained by setting all incomplete data than can be considered a yes (no), as yes (no). The complete case estimate for $\theta$ is based on the subjects answering all three questions and the available case estimate is based on the subjects answering the two questions of interest here. It is noteworthy that both estimates fall outside the pessimistic–optimistic interval and should be disregarded, since these seemingly straightforward estimators do not take the decision to treat absences as no's into account and thus discard available information. This confirms that care should be taken with the simple methods and a transition to MAR or more elaborate methods may be in place. Rubin, Stern, and Vehovar (1995) considered two MAR models, also reported in Table 6, the first one based on the two questions of direct interest only, the second one using all three. Finally, they considered a single MNAR model, based on the assumption that missingness on a question depends on the answer to that question but not on the other questions. Rubin, Stern, and Vehovar (1995) concluded, owing to the proximity of the MAR analysis to the plebiscite value, that MAR in this and similar cases may be considered a plausible assumption.

Molenberghs, Kenward, and Goetghebeur (2001) and Molenberghs *et al* (2006) fitted the BRD models and Table 7 summarizes the results. BRD1 produces $\widehat{\theta} = 0.892$, exactly the same as the first MAR estimate obtained by Rubin, Stern, and Vehovar (1995). This does not come as a surprise, since both models assume MAR and use information from the two main questions. A graphical representation of the original analyses and the BRD models combined is given in Figure 2.

## 3 Complexity of Model Selection and Assessment With Incomplete Data

Model selection and assessment are well established components of statistical analysis, whether in cross-sectional or correlated settings, including multivariate, longitudinal, and clustered data. There are several strands of intuition surrounding model selection and assessment. First, it is researchers' common understanding that "observed$\simeq$expected" for a well fitting model, which is usually understood to imply that observed and fitted profiles ought to be sufficiently similar in a longitudinal study, or observed and fitted counts in contingency tables, etc. Second, for the special case of samples from univariate or multivariate distributions, the estimators for the mean vector and the variance-covariance matrix are independent, both in a small-sample as well as in an asymptotic sense. Third, in the same situation, the least squares and maximum likelihood estimators are identical as far as mean parameters are concerned, and asymptotically equal for covariance parameters. Fourth, in a likelihood-based context, deviances and related information criteria are considered useful and practical tools for model assessment. Fifth, saturated models are uniquely defined and at the top of the model hierarchy. For contingency tables, such a saturated model is one which exactly reproduces the observed counts.

It is extremely important to realize that the five points of intuition are based on our experience with well-balanced designs and complete sets of data. We will now illustrate each of them, grouped into three categories, by means of the running examples and by general considerations.

### 3.1 The "observed$\simeq$expected" Relationship

Figure 3 shows the observed and fitted mean structures for Models 1, 2, 3, and 7, fitted to the complete version of the growth dataset, as reported in Section 2.1. Note that observed and fitted means coincide for Model 1. This is in line with general theory, since the model saturates the group by time mean structure *and*, in addition, the data are balanced. While, for the incomplete version of the data, direct likelihood nicely recovers Model 7, observed and expected means do not coincide anymore, not even under Model 1 where the group by age mean structure is saturated (see Figure 4). It is important the discrepancy is seen for the mean at age 10, the only one for which there is missingness.

### 3.2 The Mean–variance Relationship in a Normal Distribution

Let us consider Table 8 to obtain insight into the effect of the variance-covariance structure on the mean structure. We retain an unstructured group by age mean structure, and pair it with three covariance structures. Apart from an unstructured residual covariance matrix (Model 1), we also consider a CS structure (Model 7b) and an independence structure (Model 8b).

When the data are complete, the choice of covariance structure is immaterial for the point estimates, whereas the choice is crucial when data are incomplete. Next to the over-correction of Model 1 at age 10, Model 7b exhibits quite acceptable behavior, but Model 8b coincides with and hence is as bad as CC at age 10.

Figure 5 presents the mean fit associated with all three models, for both sexes. Whereas the models, fitted to the complete data, would all simply pass through the large bullets and diamonds, the differences clearly emerge from the line profiles when the models are fitted to the incomplete data.

## 3.3 The Least Squares–Maximum Likelihood Difference

Let us now turn to the difference between ordinary least squares and maximum likelihood. This is a different issue but, as we will see in what follows, it is closely related to the previous two issues.

In Table 4 we noticed that the maximum likelihood based estimates for the incomplete data differ from the OLS estimates (MANOVA and ANOVA per time point). Thus, while least-squares regression and normal distribution based regression produce the same point estimator and asymptotically the same precision estimator *for balanced and complete data*, this is no longer true in the incomplete data setting. The former method is frequentist in nature, the second one likelihood based. We will illustrate this result for a simple, bivariate normal population with missingness in Section 3.3.1. An analogous result for an incomplete contingency table will be derived in Section 3.3.2.

### 3.3.1 A Bivariate Normal Population

Consider a bivariate normal population:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right), \tag{3}$$

out of which $i = 1, \ldots, N$ subjects are sampled, each of which are supposed to provide $j = 1, 2$ measurements. Assume further that $d$ subjects complete the study and $N - d$ drop out after the first measurement occasion.

In a frequentist available case method, the parameters in (3) are estimated using the available information (Little and Rubin 2002, Verbeke and Molenberghs 2000), i.e., least squares is used. This implies $\mu_1$ and $\sigma_1^2$ would be estimated using all $N$ subjects, whereas only the remaining $d$ contribute to the other three parameters. For the mean parameters, this produces:

$$\widehat{\mu_1} = \frac{1}{N}\sum_{i=1}^{N} y_{i1}, \tag{4}$$

$$\widetilde{\mu_2} = \frac{1}{d}\sum_{i=1}^{d} y_{i2}, \tag{5}$$

To obtain an explicit expression for the likelihood-based estimators, along the lines of Little and Rubin (2002), observe that the conditional density of the second outcome given the first one, based on (3), can be written as:

$$Y_{i2}|y_{i1} \sim N(\beta_0 + \beta_1 y_{i1}, \sigma^2_{2|1}), \tag{6}$$

where

$$\left\{\begin{array}{rcl} \beta_1 & = & \rho\frac{\sigma_2}{\sigma_1}, \\ \beta_0 & = & \mu_2 - \beta_1\mu_1 = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1, \\ \sigma^2_{2|1} & = & \sigma^2_2(1-\rho^2), \\ \rho & = & \frac{\sigma_{12}}{\sigma_1\sigma_2}. \end{array}\right.$$

Now, the MLE for the first mean coincides with (4), underscoring the fact that OLS and ML provide the same point estimators, when the data are complete. For the second mean, however, we now obtain:

$$\widehat{\mu_2} = \frac{1}{N}\left\{\sum_{i=1}^{d} y_{i2} + \sum_{i=d+1}^{N}\left[\overline{y}_2 + \widehat{\beta_1}(y_{i1} - \overline{y}_1)\right]\right\}. \tag{7}$$

Here, $\overline{y}_1$ is the mean of the measurements at the first occasion among the completers. Several observations can be made. First, under MCAR, the completers and dropouts have equal distributions at the first occasion, and hence the correction term has expectation zero, rendering, again, the frequentist (least squares) and likelihood methods equivalent, *even though they do not produce exactly the same point estimator*. Second, when there is no correlation between the first and second measurements, the regression coefficient $\beta_1 = 0$, and hence there is no correction neither.

Let us assess the implications for the issues raised above, especially in the context of the orthodontic growth data. The likelihood takes the expectation into account of the missing measurements, given the observed ones. In our data, this only occurs at the age of 10. Comparing the small (all children) with the large (remaining children) bullet and diamonds, it is clear that those remaining on study have larger measurements than those removed. The direct-likelihood correction has produced estimates at the age of 10 that are situated below the observed means. Obviously, the likelihood tends to over-correct in this case. The reason for this is that the estimated correlation between the ages 8 and 10 is substantially larger than the correlation between ages 10 and 12. Such variability is not unexpected in relatively small samples. Hence, a careful reflection on the variance-covariance structure is much more important here than when data are complete and balanced. We return to these points in the next section. There are also important consequences for model checking, since the difference between observed and expected quantities can be a function of relatively poor fit *and* the adjustment of the estimates for missingness not of the MCAR type. We return to this in Section 4.

Additionally, the coefficient $\beta_1$ depends on the variance components $\sigma_1^2$, $\sigma_{12}$, and $\sigma_2^2$. This implies a misspecified variance structure may lead to bias in $\widehat{\mu_2}$. Thus, the well-known independence between the distributions of the estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in multivariate normal population holds, once again, only when the data are balanced and complete.

The adequate performance of Model 7b owes to the fact that the expected mean of a missing age 10 measurements gives equal weight to all surrounding measurements, rather than overweighting the age 8 measurement due to an accidentally high correlation. The zero correlations in Model 8b, do not allow for such a correction and hence the information that the ages 8, 12, and 14 measurements for the incomplete profiles are relatively low is wasted.

### 3.3.2 An Incomplete Contingency Table

Analogous to the incomplete bivariate normal sample of the previous section, it is insightful to consider an incomplete $2 \times 2$ contingency table:

$$\begin{array}{|c|c|}\hline Z_{1,11} & Z_{1,12} \\ \hline Z_{1,21} & Z_{1,22} \\ \hline \end{array} \qquad \begin{array}{|c|}\hline Z_{0,1} \\ \hline Z_{0,2} \\ \hline \end{array}, \tag{8}$$

where $Z_{r,jk}$ refers to the number of subjects in the completers $(r = 1)$ and dropout $(r = 0)$ groups, respectively, with response profile $(j, k)$. Since for the dropouts only the first outcome is observed, only summaries $Z_{r=0,j}$ are observable. Using all available data, the probability of success at the first time is estimated as:

$$\widehat{\pi_1} = \frac{Z_{1,1+} + Z_{0,1+}}{N}, \tag{9}$$

where '+' instead of a subscript refers to summing over the corresponding subscript. When the available cases only are used, the estimator for the success probability at the second time is:

$$\widetilde{\pi_2} = \frac{Z_{1,+1}}{d}. \tag{10}$$

Once again, information from the incomplete subjects is not used. It is easy to show that the MLE under MAR, i.e., ignorability, equals:

$$\widehat{\pi_2} = \frac{Z_{1,+1} \ + \ Z_{0,1} \cdot \frac{Z_{1,11}}{Z_{1,1+}} \ + \ Z_{0,2} \cdot \frac{Z_{1,21}}{Z_{1,2+}}}{N}. \tag{11}$$

The second and third terms in (11) result from splitting the partially classified counts according to the corresponding conditional distributions in the first table.

### 3.4 Deviances and Saturated Models

Revisiting Table 7, we observe that a deviance comparison between BRD1 and any of BRD2–5 and of the latter with BRD6–9 shows the earlier models suffer from a poor fit. Thus, effectively, we are left with BRD6–9 as candidate models for the SPO data. However, all four models produce exactly the same likelihood at maximum. This is not surprising, since the models contain eight parameters, equal to the number of degrees of freedom in the *observed* data. Nevertheless, the estimates for $\theta$ differ between these four models. The reason is that $\theta$ is a function, not only of the model fit to the observed data, but of the model's *prediction* for the unobserved data, given what has been observed. Thus, model fit and the concept of saturation can be seen either as relative to the observed data, or relative to the complete data, observed and unobserved simultaneously. This, again, poses specific challenges for model selection and assessment of model fit.

All models BRD6–9 being of the MNAR type, it is tempting to conclude that all evidence points to MNAR as the most plausible missing data mechanism. Nothwithstanding this observation, one cannot even so much as formally exclude MAR. Indeed, Molenberghs *et al* (2006) have shown that for every MNAR model considered, there is an associated MAR "bodyguard", a model reproducing the same fit as the original MNAR model, but predicting the unobserved data given the observed ones consistent with MAR. Formal derivations are given in Molenberghs *et al* (2006). The corresponding estimates for the proportion $\theta$ in favor of independence are presented in the last column of Table 7. Let us informally study the relationship and its implications by means of models BRD1, BRD2, BRD7, and BRD9, fitted to the SPO data. BRD1 assumes MCAR, all others MNAR. Only BRD7 and BRD9 saturate the observed-data degrees of freedom. The incomplete data as observed, as predicted by each of the four models, and as predicted by these four models' MAR counterparts, are displayed in Table 9. The corresponding predictions of the hypothetical, complete data are presented in Table 10. The fits of models BRD7, BRD9, and their MAR counterparts, coincide with the observed data. As follows from Molenberghs *et al* (2006) every model produces exactly the same fit as does its MAR counterpart; hence, this is seen for all four models. Since BRD1 is MCAR and hence MAR to begin with, it is the only coinciding with its MAR counterpart, since indeed BRD1$\equiv$BRD1(MAR). Further, while BRD7 and BRD9 produce a different fit to the complete data, BRD7(MAR) and BRD9(MAR) coincide. This is because the fits of BRD7 and BRD9 coincide with respect to their fit to the observed data; because they are saturated, they coincide as such with the incomplete, observed data.

An important observation for model assessment and selection is that the five models BRD6, BRD7, BRD8, BRD9, and BRD6(MAR)$\equiv$BRD7(MAR)$\equiv$BRD8(MAR)$\equiv$BRD9 at the same time saturate the observed data degrees of freedom and exhibit a dramatically different prediction of the full data. Thus, five perfectly fitting models produce five different estimates for the proportion in favor of independence: 0.741, 0.764, 0.867, 0.819, and 0.892.

This problem needs careful consideration and it is very clear that there are instances where a model cannot be selected merely on classical model comparison and selection tools.

Additional problems can occur, such as predicted complete tables with negative counts, as reported by Baker, Rosenberger, and DerSimonian (1992), Molenberghs *et al* (1999), and Molenberghs and Kenward (2007).

# 4 Model Selection and Assessment with Incomplete Data

The five issues laid out at the start of Section 3 and illustrated using both examples essentially originate from the fact that, when fitting models to incomplete data, one needs to manage two aspects rather than a single one, as schematically represented in Figure 6: the contrast between data and model is supplemented with a second contrast between their complete and incomplete versions.

Ideally, we would want to consider the situation depicted in Figure 6(b), where the comparison is fully made at the complete level. Since the complete data are, by definition, beyond reach, it is tempting but dangerous to settle for the situation in Figure 6(c). This would happen when we would conclude Model 1 fit poorly to the orthodontic growth data, as elucidated by Figure 4. Such a conclusion would ignore that the model fit is at the complete-data level, accounting for 16 boys and 11 girls at the age of 10, whereas the data only represent the remaining 11 boys and 7 girls at the age of 10. Thus, a fair model assessment should be confined to the situations laid out in Figure 6(b) and (d) only. We will start out by the simpler (d) and then return to (b).

Assessing whether Model 1 fits the incomplete version of the growth dataset well can be done by comparing the observed means at the age of 10 to its prediction by the model. This implies we have to confine model fit to those children actually observed at the age of 10.

Turning to the analysis of the Slovenian public opinion survey, the principle behind Figure 6(d) would lead to the conclusion that the five models BRD6, BRD7, BRD8, BRD9, and BRD6(MAR)≡ BRD7(MAR)≡BRD8(MAR)≡BRD9 perfectly fit the observed data, as can be seen in Figure 9 (first panel). As we stated earlier, though, the models are drastically different in their complete-data level fit (Figure 10) and the corresponding estimates of the proportion in favor of independence, which ranges over $[0.74; 0.89]$. This points to the need for supplementing model assessment, even when done in the preferable situation Figure 6(d), with a form of sensitivity analysis.

In conclusion, there are *two* important aspects in selection and assessment when data are incomplete. First, the model needs to fit the *observed* data well. This aspect alone is already quite a bit more complicated than in the complete/balanced case as shown in Section 3. We will expand on this first aspect in Section 4.1. Second, sensitivity analysis is advisable to assess in how far the model selected and conclusions reached are sensitive to the explicit or implicit assumptions a model makes about the incomplete data given the observed ones because such assumptions typically have an impact on the inferences of interest. This aspect is elaborated upon in Section 4.2.

## 4.1 Model Fit to Observed Data

As stated before, model fit to the observed data can be done by means of either what we will label Scenario I, as laid out in Figure 6(b), or by means of Scenario II of Figure 6(d).

Under Scenario I, we conclude BRD6–9 or their MAR counterpart fit perfectly. There is nothing wrong with such a conclusion, as long as we realize *there is more than one model* with this very same property, while at the same time they lead to different substantive conclusions. If one would have started with a single one from amongst these models without considering any of the others there is a real danger when the conclusions are based on that particular model only. For example, if one would so choose BRD9, the conclusion would be that $\widehat{\theta} = 0.867$ with 95% confidence interval $[0.851; 0, 884]$. Ignoring the other perfectly fitting models does not make sense, unless there are very strong substantive reasons to do so.

Turning to the orthodontic growth data, considering the fit of Model 1 to the data has some interesting ramifications. When the OLS fit is considered, only valid under MCAR, one would conclude there is a perfect fit to the observed means, also at the age of 10. The fit using ML would apparently show a discrepancy, since the observed mean refers to a reduced sample size while the fitted mean, similar to (7), is based on the entire design.

These considerations suggest that we consider the fit of a model to an incomplete set of data requires caution and perhaps extension and/or modification of the classical model assessment paradigms. In particular, it is of interest to consider assessment under Scenario II.

Gelman *et al* (2005) proposed a Scenario II method. The essence of their approach is as follows. First, a model, saturated or non-saturated, is fitted to the observed data. Under the fitted model, and assuming ignorable missingness, datasets simulated from the fitted model should 'look similar' to the actual data. Therefore, multiple sets of data are sampled from the fitted model, and compared to the dataset at hand. Because what one actually observes consists of, not only the actually observed outcome data, but also realizations of the missingness process, comparison with the simulated data would also require simulation from, hence full specification of, the missingness process. This added complexity is avoided by augmenting the observed outcomes with imputations drawn from the fitted model, conditional on the observed responses, and by comparing the so-obtained completed dataset with the multiple versions of simulated complete datasets. Such a comparison will usually be based on relevant summary characteristics such as time-specific averages or standard deviations. As suggested by Gelman *et al* (2005), this so-called data-augmentation step could be done multiple times, along multiple-imputation ideas from Rubin (1987). However, in cases with a limited amount of missing observations, the between-imputation variability will be far less important than the variability observed between multiple simulated datasets. This is in contrast to other contexts to which the technique of Gelman *et al* (2005) has been applied, e.g., situations where latent unobservable variables are treated as 'missing'.

Let us first apply the method to the orthodontic growth data. The first model considered assumes a saturated mean structure, as in Model 1, with a compound-symmetric covariance structure (Model 1a). Twenty datasets are simulated from the fitted model, and time-specific sample averages are compared to the averages obtained from augmenting the observed data based on the fitted model. The results are shown in the top panel of Figure 7. The sample average at age 10, for the girls, is relatively low compared to what would be expected under the fitted model. Since the mean structure is saturated, this may indicate lack of fit of the covariance structure. We therefore extend the model by allowing for sex-specific covariance structures (Model 1b). The results under this new model are presented in the bottom panel of Figure 7. The observed data are now less extreme compared to what is expected under the fitted model. Formal comparison of the two models, based on a likelihood ratio test, indeed rejects the first model in favor of the second one ($p = 0.0003$), with much more between-subject variability for the girls than for the boys, while the opposite is true for the within-subject variability.

Let us now turn to the SPO data. In such a contingency table case, the above approach can be simplified to comparing the model fit to the complete data, such as presented in Table 9, with their counterpart obtained from extending the observed, incomplete data to their complete counterpart by means of the fitted model. Here, we have to distinguish between saturated and non-saturated models. For saturated models, such as BRD6–9 and their MAR counterparts, this is simply the same table as the model fit and again, all models are seen to fit perfectly. Of course, this statement needs further qualification. It still merely means that these models fit the *incomplete* data perfectly, while each one of them tells a different, unverifiable story about the unobserved data given the observed ones. In contrast, for the non-saturated models, such as BRD1–5 and their MAR counterparts, a so-completed table is different from the fitted one. To illustrate this, the completed tables are presented in Table 11, for the same set of models as in Tables 9 and 10.

A number of noteworthy observations can be made. First, BRD1≡BRD1(MAR) exhibit the poorest fit (i.e., the largest discrepancies between this completed table and the model fit as presented in Table 10), with an intermediate quality fit for a model with 7 degrees of freedom, such as BRD2, and a perfect fit for BRD7, BRD9, and their MAR counterparts. Second, compare the data completed using BRD1 (Table 11) to the fit of BRD1 (Table 10): the data for the group of completers is evidently equal to the original data (Table 9) since here no completion is necessary; the complete data for the subjects without observations is entirely equal to the model fit (Table 10), since here there are no data to start from; the complete data for the two partially classified tables takes a position in between and hence is not exactly equal to the model fit. Third, note that the above statement is in need of amendment for BRD2 and BRD2(MAR). Now, the first subtable of partially classified subjects exhibits an exact match between completed data and model fit, while this is not true for the second subtable. The reason is that BRD2 allows missingness on the second question to depend on the first one, leading to saturation of the first subtable, whereas missingness on the first question is independent of one's opinion on either question.

While the method is elegant and gives us a handle regarding the quality of the model fit to the incomplete data while contemplating the completed data and the full model fit, the method is unable to distinguish between the saturated models BRD6–9 and the MAR counterpart, as any method would. This phenomenon points to the need for sensitivity analysis, a topic taken up next.

## 4.2 Sensitivity Analysis

In the previous section, we have seen how one can proceed to assess model fit, either under Scenario I or using Scenario II. It is important to reiterate this comprises the fit to the observed data only, and strictly makes no statement about the model in as far as it describes, or predicts, the unobserved given the observed data. To address the latter issue, a variety of sensitivity analysis routes have been proposed. One could informally define a sensitivity analysis as a way of exploring the impact of a model and/or selected observations on the inferences made when data are incomplete.

For example, Verbeke *et al* (2001), Thijs, Molenberghs, and Verbeke (2000), Molenberghs *et al* (2001), Van Steen *et al* (2001), and Jansen *et al* (2003) advocated the use of local influence based methods for sensitivity analysis purposes. The essence of the method is that (i) a subject-specific perturbation is added to the model, e.g., by replacing the parameter describing MNAR missingness in the model by Diggle and Kenward (1994) with a subject-specific perturbation:

$$\text{logit}[\text{P}(\text{dropout at occasion } j|y_{i,j-1}, y_{i,j})] = \psi_0 + \psi_1 y_{i,j-1} + \omega_i \psi_2 y_{ij} \tag{12}$$

(ii) then observing that $\omega_i \equiv 0$ corresponds to MAR, and (iii) finally studying the impact of small perturbations of $\omega_i$ around zero. (Indeed, a model like (12) is necessary, since for an MNAR model, not only the measurements need to be modeled (e.g., using a linear mixed model), also the dropout mechanism needs to be modeled as a function of the measurements and, in some cases, covariates.) Technically, this is done using differential geometry methods. In a variety of examples, the above authors showed that one or a few observations are sometimes able to drive the conclusions about the missing data mechanism. Details can be found in the aforementioned references, as well as in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005). We applied the method to the orthodontic growth data, assuming either Model 1 or Model 7 of Table 2. The results are qualitatively the same and we present the Model 1 results only. Subjects #3 (girl), and #13, #23, and #27 (boys) come out as very influential. In addition, some influence is seen for #6 and #9 (girls), and #16 (boy). As can be seen from Figure 8, all of these are incomplete, which is different from other applications of the method. Of course, all but one of these are positioned relatively low, and one cannot conclude definitively whether either their incompleteness status or the location of their profile is determining their influence. The influence measure informally described above and denoted by $C_i$ is presented in Figure 9. Even though the $C_i$ measure exhibits very high peaks, removing the highly influential subjects does not alter the substantive conclusions.

Molenberghs, Kenward, and Goetghebeur (2001) and Kenward, Goetghebeur, and Molenberghs (2001) suggested the use of so-called *regions of ignorance*, combining uncertainty owing to finite sampling with uncertainty resulting from incompleteness. Broadly speaking, they consider overspecified models which then produce non-unique solutions of the likelihood equations. For a single (vector) parameter, the resulting solution is called the interval (region) of ignorance. When uncertainty due to finite sampling is added, a wider interval (region) of uncertainty is obtained. For the SPO data, this comes down to considering models with nine or more degrees of freedom.

The estimated intervals of ignorance and intervals of uncertainty are shown in Table 12, while a graphical representation of the YES votes is given in Figure 10. Model 10 is defined as $(\alpha_k, \beta_{jk})$ with

$$\beta_{jk} = \beta_0 + \beta_j + \beta_k, \qquad (13)$$

while Model 11 assumes $(\alpha_{jk}, \beta_j)$ and uses

$$\alpha_{jk} = \alpha_0 + \alpha_j + \alpha_k, \qquad (14)$$

Finally, Model 12 is defined as $(\alpha_{jk}, \beta_{jk})$, a combination of both (13) and (14). Model 10 shows an interval of ignorance which is very close to $[0.741, 0.892]$, the range produced by the models BRD1–BRD9, while Model 11 is somewhat sharper and just fails to cover the plebiscite value. However, it should be noted that the corresponding intervals of uncertainty contain the true value.

Interestingly, Model 12 virtually coincides with the non-parametric range even though it does not saturate the complete data degrees of freedom. To do so, not 2 but in fact 7 sensitivity parameters would have to be included. Thus, it appears that a relatively simple sensitivity analysis is sufficient to increase the insight in the information provided by the incomplete data about the proportion of valid YES votes.

## 5 Concluding Remarks

In this paper, we have illustrated the complexities arising when fitting models to incomplete data. By means of two case studies, the continuous longitudinal orthodontic growth data and the discrete Slovenian Public Opinion Survey data, five generic issues were brought to the forefront: (i) the classical relationship between observed and expected features is convoluted since one observes the data only partially while the model describes all data; (ii) the independence of mean and variance parameters in a (multivariate) normal is lost, implying increased sensitivity, even under MAR; (iii) also the well-known agreement between the (frequentist) OLS and maximum likelihood estimation methods for normal models is lost, as soon as the missing data mechanism is not of the MCAR type, with related results holding in the non-normal case; (iv) in a likelihood-based context, deviances and related information criteria cannot be used in the same way as with complete data since they provide no information about a model's prediction of the unobserved data and, in particular, (v) several models may saturate the observed-data degrees of freedom, while providing a different fit to the complete data, i.e., they only coincide in as far as they describe the observed data; as a consequency, different inferences may result from different saturated models.

Based on these considerations it is argued that model assessment should always proceed in two steps. In the first step, the fit of a model to the *observed* data should be assessed carefully, while in the second step the sensitivity of the conclusions to the *unobserved data given the observed data* should be addressed. In the first step, one should ensure that the required assessment be done under one of two allowable scenarios, as represented by Figures 6(b) and (d), thereby carefully avoiding the scenario of Figure 6(c), where the model at the complete data level is compared to the incomplete data; apples and oranges as it were. The method proposed by Gelman *et al* (2005) offers a convenient route to model assessment.

## References

AFIFI, A. and ELASHOFF, R. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association* **61** 595–604.

BAKER, S.G., ROSENBERGER, W.F., and DERSIMONIAN, R. (1992). Closed-form estimates for missing counts in two-way contingency tables, *Statistics in Medicine* **11** 643–657.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39** 1–38.

DIGGLE, P.J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics* **45** 1255–1258.

GELMAN, A., VAN MECHELEN, I., VERBEKE, G., HEITJAN, D.F., and MEULDERS, M. (2005). Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* **61** 74–85.

HARTLEY, H.O. and HOCKING, R. (1971). The analysis of incomplete data. *Biometrics* **27** 7783–808.

JANSEN, I., BEUNCKENS, C., MOLENBERGHS, G., VERBEKE, G., and MALLINCKRODT, C. (2006a). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science* **21** 52–69.

JANSEN, I., HENS, N., MOLENBERGHS, G., AERTS, M., VERBEKE, G., and KENWARD, M.G. (2006b). The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis* **50** 830–858.

JANSEN, I., MOLENBERGHS, G., AERTS, M., THIJS, H., and VAN STEEN, K. (2003). A Local influence approach applied to binary data from a psychiatric study. *Biometrics* **59** 410–419.

JENNRICH, R.I. and SCHLUCHTER, M.D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics* **42** 805–820.

KENWARD, M.G, GOETGHEBEUR, E.J.T., and MOLENBERGHS, G. (2001). Sensitivity analysis of incomplete categorical data. *Statistical Modelling* **1** 31–48.

LITTLE, R.J.A. and RUBIN, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: John Wiley & Sons.

MALLINCKRODT, C.H., CARROLL, R.J., DEBROTA, D.J., DUBE, S., MOLENBERGHS, G., POTTER, W.Z., SANGER, T.D., and TOLLEFSON, G.D. (2003a). Assessing and interpreting treatment effects in longitudinal clinical trials with subject dropout. *Biological Psychiatry* **53** 754–760.

MALLINCKRODT, C.H., SCOTT CLARK, W., CARROLL, R.J., and MOLENBERGHS, G. (2003b). Assessing response profiles from incomplete longitudinal clinical trial data with subject dropout under regulatory conditions. *Journal of Biopharmaceutical Statistics* **13** 179–190.

MOLENBERGHS, G., BEUNCKENS, C., SOTTO, C., and KENWARD, M.G. (2006). Every missing not at random model has got a missing at random bodyguard. *Submitted for publication*.

MOLENBERGHS, G., GOETGHEBEUR, E.J.T., LIPSITZ, S.R., KENWARD, M.G. (1999). Non-random missingness in categorical data: strengths and limitations. *The American Statistician* **53** 110–118.

MOLENBERGHS, G., KENWARD, M.G., and GOETGHEBEUR, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics* **50** 15–29.

MOLENBERGHS, G., THIJS, H., JANSEN, I., BEUNCKENS, C., KENWARD, M.G., MALLINCKRODT, C., and CARROLL, R.J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5** 445-464.

MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

MOLENBERGHS, G., VERBEKE, G., THIJS, H., LESAFFRE, E., and KENWARD, M.G. (2001). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process, *Computational Statistics and Data Analysis* **37** 93–113.

MOLENBERGHS, G. and KENWARD, M.G. (2007). *Handling Incomplete Data From Clinical Studies.* New York: Wiley.

POTTHOFF, R.F. and ROY, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51** 313–326.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

THIJS, H., MOLENBERGHS, G., and VERBEKE, G. (2000). The milk protein trial: influence analysis of the dropout process. *Biometrical Journal* **42** 617–646.

VAN STEEN, K., MOLENBERGHS, G., VERBEKE, G., and THIJS, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal* **1** 125–142.

VERBEKE, G. and MOLENBERGHS, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach.* Lecture Notes in Statistics 126. New York: Springer.

VERBEKE, G. and MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

VERBEKE, G., MOLENBERGHS, G., THIJS, H., LESAFFRE, E., and KENWARD, M.G. (2001). Sensitivity analysis for non-random dropout: a local influence approach, *Biometrics* **57** 7–14.

Table 1: *The Orthodontic Growth Data. Data for 11 girls and 16 boys. Measurements marked with * were deleted by Little and Rubin (2002). Original source: Pothoff and Roy (1964), Jennrich and Schluchter (1986).*

| | Age (in years) | | | | | Age (in years) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Girl | 8 | 10 | 12 | 14 | Boy | 8 | 10 | 12 | 14 |
| 1 | 21.0 | 20.0 | 21.5 | 23.0 | 1 | 26.0 | 25.0 | 29.0 | 31.0 |
| 2 | 21.0 | 21.5 | 24.0 | 25.5 | 2 | 21.5 | 22.5* | 23.0 | 26.5 |
| 3 | 20.5 | 24.0* | 24.5 | 26.0 | 3 | 23.0 | 22.5 | 24.0 | 27.5 |
| 4 | 23.5 | 24.5 | 25.0 | 26.5 | 4 | 25.5 | 27.5 | 26.5 | 27.0 |
| 5 | 21.5 | 23.0 | 22.5 | 23.5 | 5 | 20.0 | 23.5* | 22.5 | 26.0 |
| 6 | 20.0 | 21.0* | 21.0 | 22.5 | 6 | 24.5 | 25.5 | 27.0 | 28.5 |
| 7 | 21.5 | 22.5 | 23.0 | 25.0 | 7 | 22.0 | 22.0 | 24.5 | 26.5 |
| 8 | 23.0 | 23.0 | 23.5 | 24.0 | 8 | 24.0 | 21.5 | 24.5 | 25.5 |
| 9 | 20.0 | 21.0* | 22.0 | 21.5 | 9 | 23.0 | 20.5 | 31.0 | 26.0 |
| 10 | 16.5 | 19.0* | 19.0 | 19.5 | 10 | 27.5 | 28.0 | 31.0 | 31.5 |
| 11 | 24.5 | 25.0 | 28.0 | 28.0 | 11 | 23.0 | 23.0 | 23.5 | 25.0 |
| | | | | | 12 | 21.5 | 23.5* | 24.0 | 28.0 |
| | | | | | 13 | 17.0 | 24.5* | 26.0 | 29.5 |
| | | | | | 14 | 22.5 | 25.5 | 25.5 | 26.0 |
| | | | | | 15 | 23.0 | 24.5 | 26.0 | 30.0 |
| | | | | | 16 | 22.0 | 21.5* | 23.5 | 25.0 |

Table 2: *The Orthodontic Growth Data. Original and trimmed data set. Model fit summary. ('#par': number of model parameters; $-2\ell$: minus twice log-likelihood; Ref: reference model for likelihood ratio test).*

| | | | | | Original data | | Trimmed data | |
|---|---|---|---|---|---|---|---|---|
| Model | Mean | Covar. | #par | Ref | $-2\ell$ | $p$-value | $-2\ell$ | $p$-value |
| 1 | unstr. | unstr. | 18 | | 416.5 | | 386.96 | |
| 2 | $\neq$ slopes | unstr. | 14 | 1 | 419.5 | 0.563 | 393.29 | 0.176 |
| 3 | $=$ slopes | unstr. | 13 | 2 | 426.2 | 0.010 | 397.40 | 0.043 |
| 4 | $\neq$ slopes | Toepl. | 8 | 2 | 424.6 | 0.523 | 398.03 | 0.577 |
| 5 | $\neq$ slopes | AR(1) | 6 | 2 | 440.7 | 0.007 | 409.52 | 0.034 |
| 6 | $\neq$ slopes | RI+RS | 8 | 2 | 427.8 | 0.215 | 400.45 | 0.306 |
| 7 | $\neq$ slopes | CS (RI) | 6 | 6 | 428.6 | 0.510 | 401.31 | 0.502 |
| 8 | $\neq$ slopes | simple | 5 | 7 | 478.2 | <0.001 | 441.58 | <0.001 |

Table 3: *The Orthodontic Growth Data. Finally selected model under a number of simple missing data handling mechanisms. (par: # parameters; a suffix 'a' to a model number refers to a variation to one of the models in Table 2)*

| Method | Model | Mean | Covar | par |
|---|---|---|---|---|
| Complete case | 7a | $=$ slopes | CS | 5 |
| LOCF | 2a | quadratic | unstructured | 16 |
| Unconditional mean | 7a | $=$ slopes | CS | 5 |
| Conditional mean | 1 | unstructured | unstructured | 18 |

Table 4: *The Orthodontic Growth Data. Likelihood, MANOVA, and ANOVA analyses for the original data and the trimmed data (observed, CC, and LOCF). Means for boys at ages 8 and 10 are displayed.*

| Principle | Method | Boys at Age 8 | Boys at Age 10 |
|---|---|---|---|
| *Original* | *ML* | *22.88 (0.56)* | *23.81 (0.49)* |
| | *REML ≡ MANOVA* | *22.88 (0.58)* | *23.81 (0.51)* |
| | *ANOVA per time* | *22.88 (0.61)* | *23.81 (0.53)* |
| Observed | ML | 22.88 (0.56) | 23.17 (0.68) |
| | REML | 22.88 (0.58) | 23.17 (0.71) |
| | MANOVA | 24.00 (0.48) | 24.14 (0.66) |
| | ANOVA per time | 22.88 (0.61) | 24.14 (0.74) |
| CC | ML | 24.00 (0.45) | 24.14 (0.62) |
| | REML ≡ MANOVA | 24.00 (0.48) | 24.14 (0.66) |
| | ANOVA per time | 24.00 (0.51) | 24.14 (0.74) |
| LOCF | ML | 22.88 (0.56) | 22.97 (0.65) |
| | REML ≡ MANOVA | 22.88 (0.58) | 22.97 (0.68) |
| | ANOVA per time | 22.88 (0.61) | 22.97 (0.72) |

Table 5: *The Slovenian Public Opinion Survey. The* Don't Know *category is indicated by* *.

| | | Independence | | |
|---|---|---|---|---|
| Secession | Attendance | Yes | No | * |
| Yes | Yes | 1191 | 8 | 21 |
| | No | 8 | 0 | 4 |
| | * | 107 | 3 | 9 |
| No | Yes | 158 | 68 | 29 |
| | No | 7 | 14 | 3 |
| | * | 18 | 43 | 31 |
| * | Yes | 90 | 2 | 109 |
| | No | 1 | 2 | 25 |
| | * | 19 | 8 | 96 |

Table 6: *The Slovenian Public Opinion Survey. Some estimates of the proportion $\theta$ attending the plebiscite and voting for independence, as presented in Rubin, Stern, and Vehovar (1995) and Molenberghs, Kenward, and Goetghebeur (2001).*

| Estimation method | Voting in favour<br>of independence<br>$\widehat{\theta}$ |
|---|---|
| Pessimistic-optimistic bounds | [0.694;0.905] |
| Complete cases | 0.928 |
| Available cases | 0.929 |
| MAR (2 questions) | 0.892 |
| MAR (3 questions) | 0.883 |
| MAR | 0.782 |
| Plebiscite | 0.885 |

Table 7: *The Slovenian Public Opinion Survey. Analysis, restricted to the independence and attendance questions. Summaries on each of the Models BRD1–BRD9 are presented.*

| Model | Structure | d.f. | loglik | $\widehat{\theta}$ | C.I. | $\widehat{\theta}_{\text{MAR}}$ |
|---|---|---|---|---|---|---|
| BRD1 | $(\alpha, \beta)$ | 6 | -2495.29 | 0.892 | [0.878;0.906] | 0.8920 |
| BRD2 | $(\alpha, \beta_j)$ | 7 | -2467.43 | 0.884 | [0.869;0.900] | 0.8915 |
| BRD3 | $(\alpha_k, \beta)$ | 7 | -2463.10 | 0.881 | [0.866;0.897] | 0.8915 |
| BRD4 | $(\alpha, \beta_k)$ | 7 | -2467.43 | 0.765 | [0.674;0.856] | 0.8915 |
| BRD5 | $(\alpha_j, \beta)$ | 7 | -2463.10 | 0.844 | [0.806;0.882] | 0.8915 |
| BRD6 | $(\alpha_j, \beta_j)$ | 8 | -2431.06 | 0.819 | [0.788;0.849] | 0.8919 |
| BRD7 | $(\alpha_k, \beta_k)$ | 8 | -2431.06 | 0.764 | [0.697;0.832] | 0.8919 |
| BRD8 | $(\alpha_j, \beta_k)$ | 8 | -2431.06 | 0.741 | [0.657;0.826] | 0.8919 |
| BRD9 | $(\alpha_k, \beta_j)$ | 8 | -2431.06 | 0.867 | [0.851;0.884] | 0.8919 |

Table 8: *The Orthodontic Growth Data. Comparison of mean estimates for boys at ages 8 and 10, complete and incomplete data, using direct likelihood, an unstructured mean model, and various covariance models.*

| Data | Mean | Covar | Boys at Age 8 | Boys at Age 10 |
|---|---|---|---|---|
| Complete | unstr. | unstr. | 22.88 | 23.81 |
| | unstr. | CS | 22.88 | 23.81 |
| | unstr. | simple | 22.88 | 23.81 |
| Incomplete | unstr. | unstr. | 22.88 | 23.17 |
| | unstr. | CS | 22.88 | 23.52 |
| | unstr. | simple | 22.88 | 24.14 |

Table 9: *The Slovenian Public Opinion Survey. Analysis restricted to the independence and attendance questions. The observed data are shown, as well as the fit of models BRD1, BRD2, BRD7, and BRD9, and their MAR counterparts, to the observed data.*

| | | | | | |
|---|---|---|---|---|---|
| Observed data & fit of BRD7, BRD7(MAR), BRD9, and BRD9(MAR) to incomplete data | | | | | |
| 1439 | 78 | 159 | 144 | 54 | 136 |
| 16 | 16 | 32 | | | |
| Fit of BRD1 and BRD1(MAR) to incomplete data | | | | | |
| 1381.6 | 101.7 | 182.9 | 179.7 | 18.3 | 136.0 |
| 24.2 | 41.4 | 8.1 | | | |
| Fit of BRD2 and BRD2(MAR) to incomplete data | | | | | |
| 1402.2 | 108.9 | 159.0 | 181.2 | 16.8 | 136.0 |
| 15.6 | 22.3 | 32.0 | | | |

Table 10: *The Slovenian Public Opinion Survey. Analysis restricted to the independence and attendance questions. The fit of models BRD1, BRD2, BRD7, and BRD9, and their MAR counterparts, to the hypothetical complete data is shown.*

Fit of BRD1 and BRD1(MAR) to complete data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1381.6 | 101.7 | 170.4 | 12.5 | 176.6 | 13.0 | 121.3 | 9.0 |
| 24.2 | 41.4 | 3.0 | 5.1 | 3.1 | 5.3 | 2.1 | 3.6 |

Fit of BRD2 to complete data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1402.2 | 108.9 | 147.5 | 11.5 | 179.2 | 13.9 | 105.0 | 8.2 |
| 15.6 | 22.3 | 13.2 | 18.8 | 2.0 | 2.9 | 9.4 | 13.4 |

Fit of BRD2(MAR) to complete data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1402.2 | 108.9 | 147.7 | 11.3 | 177.9 | 12.5 | 121.2 | 9.3 |
| 15.6 | 22.3 | 13.3 | 18.7 | 3.3 | 4.3 | 2.3 | 3.2 |

Fit of BRD7 to complete data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 3.2 | 155.8 | 142.4 | 44.8 | 0.4 | 112.5 |
| 16 | 16 | 0.0 | 32.0 | 1.6 | 9.2 | 0.0 | 23.1 |

Fit of BRD9 to complete data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 150.8 | 8.2 | 142.4 | 44.8 | 66.8 | 21.0 |
| 16 | 16 | 16.0 | 16.0 | 1.6 | 9.2 | 7.1 | 41.1 |

Fit of BRD7(MAR) and BRD9(MAR) to complete data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 148.1 | 10.9 | 141.5 | 38.4 | 121.3 | 9.0 |
| 16 | 18 | 11.8 | 20.2 | 2.5 | 15.6 | 2.1 | 3.6 |

Table 11: *The Slovenian Public Opinion Survey. Analysis restricted to the independence and attendance questions. Completed versions of the observed data, using the fit of the modelsBRD1, BRD2, BRD7, and BRD9, and their MAR counterparts.*

Completed data using BRD1≡BRD1(MAR) fit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 148.1 | 10.9 | 141.5 | 38.4 | 121.3 | 9.0 |
| 16 | 16 | 11.9 | 20.1 | 2.5 | 15.6 | 2.1 | 3.6 |

Completed data using BRD2 fit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 147.5 | 11.5 | 142.4 | 44.7 | 105.0 | 8.2 |
| 16 | 16 | 13.2 | 18.8 | 1.6 | 9.3 | 9.4 | 13.4 |

Completed data using BRD2(MAR) fit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 147.7 | 11.3 | 141.4 | 40.2 | 121.2 | 9.3 |
| 16 | 16 | 13.3 | 18.7 | 2.6 | 13.8 | 2.3 | 3.2 |

Completed data using BRD7 fit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 3.2 | 155.8 | 142.4 | 44.8 | 0.4 | 112.5 |
| 16 | 16 | 0.0 | 32.0 | 1.6 | 9.2 | 0.0 | 23.1 |

Completed data using BRD9 fit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 150.8 | 8.2 | 142.4 | 44.8 | 66.8 | 21.0 |
| 16 | 16 | 16.0 | 16.0 | 1.6 | 9.2 | 7.1 | 41.1 |

Completed data using BRD7(MAR)≡BRD9(MAR) fit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1439 | 78 | 148.1 | 10.9 | 141.5 | 38.4 | 121.3 | 9.0 |
| 16 | 18 | 11.8 | 20.2 | 2.5 | 15.6 | 2.1 | 3.6 |

Table 12: *The Slovenian Public Opinion Survey. Intervals of ignorance and intervals of uncertainty for the proportion $\theta$ (confidence interval) attending the plebiscite following from fitting.*

| | | | $\widehat{\theta}$ | |
|---|---|---|---|---|
| Model | d.f. | loglik | II | IU |
| Model 10 | 9 | -2431.06 | [0.762;0.893] | [0.744;0.907] |
| Model 11 | 9 | -2431.06 | [0.766;0.883] | [0.715;0.920] |
| Model 12 | 10 | -2431.06 | [0.694;0.905] | |

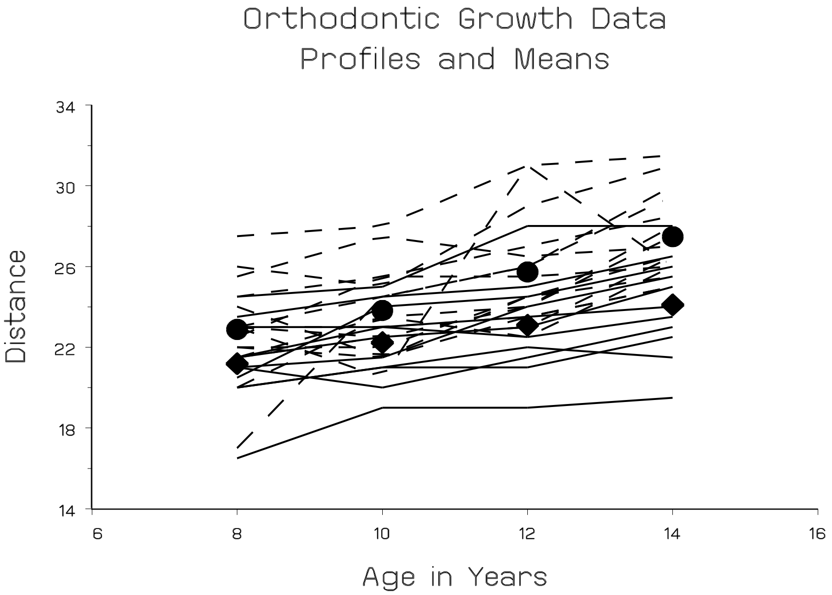


Figure 1: *The Orthodontic Growth Data. Observed profiles and group by age means. Solid lines and diamonds are for girls, dashed lines and bullets are for boys.*

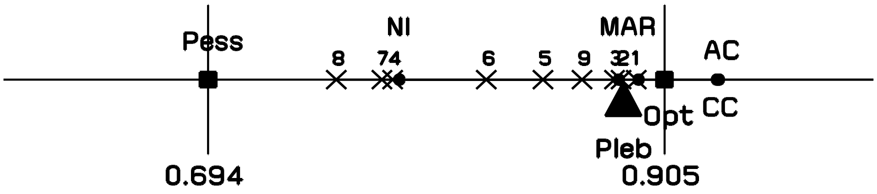


Figure 2: *The Slovenian Public Opinion Survey. Relative position for the estimates of "proportion of YES votes", based on the models considered in Rubin, Stern, and Vehovar (1995) and on the Baker, Rosenberger, and DerSimonian (1992) models. The vertical lines indicate the non-parametric pessimistic–optimistic bounds. (Pess: pessimistic boundary; Opt: optimistic boundary; MAR: Rubin* et al*'s MAR model; NI: Rubin* et al*'s MNAR model; AC: available cases; CC: complete cases; Pleb: plebiscite outcome. Numbers refer to the BRD models.)*

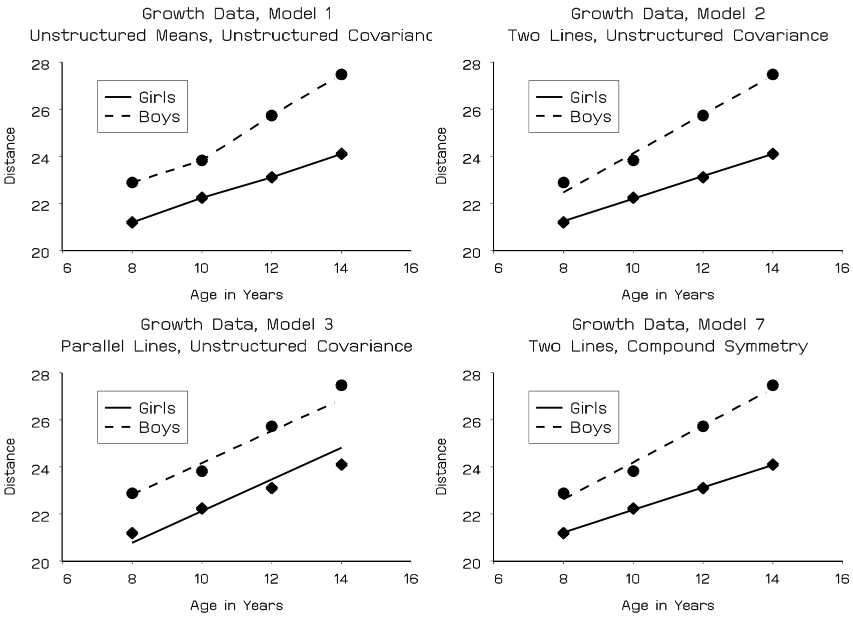


Figure 3: *The Orthodontic Growth Data. Profiles for the complete data, for a selected set of models.*

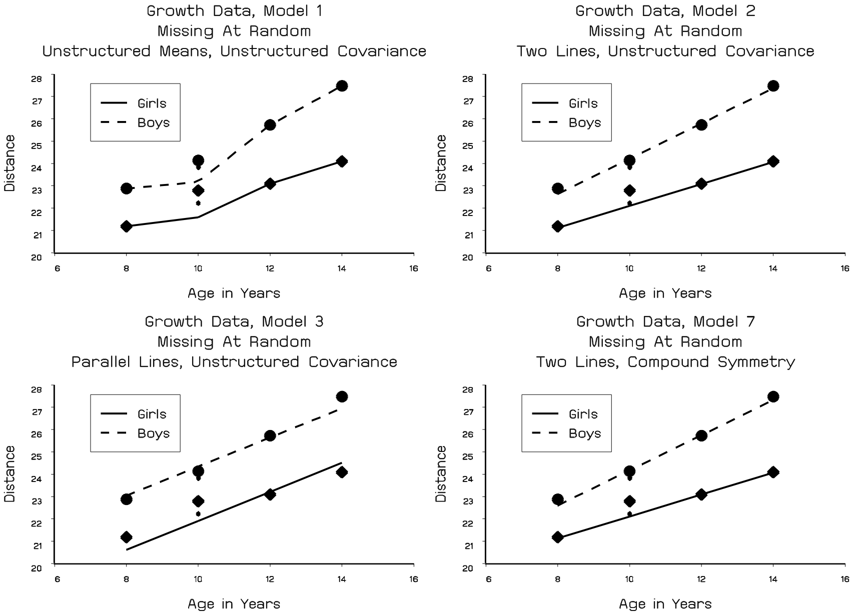


Figure 4: *The Orthodontic Growth Data. Profiles for the growth data set, from a selected set of models. MAR analysis. (The small symbols at age 10 are the observed group means for the complete data set.)*

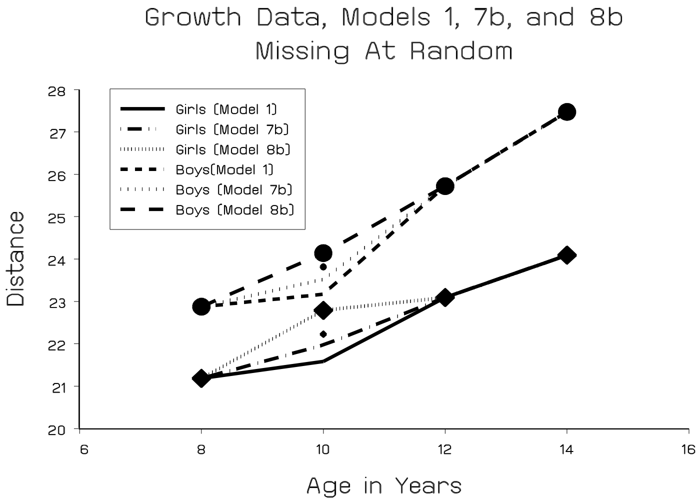


Figure 5: *The Orthodontic Growth Data. Fitted mean profiles to the incomplete data, using maximum likelihood, an unstructured mean model and unstructured (Model 1), CS (Model 7b), and independence (Model 8b) covariance structure.*

(a)

| | | **model** | |
|---|---|---|---|
| | | complete | incomplete |
| **raw** | complete | | |
| **data** | incomplete | | |

(b)

| | | **model** | |
|---|---|---|---|
| | | complete | incomplete |
| **raw** | complete | $*$ | |
| **data** | incomplete | | |

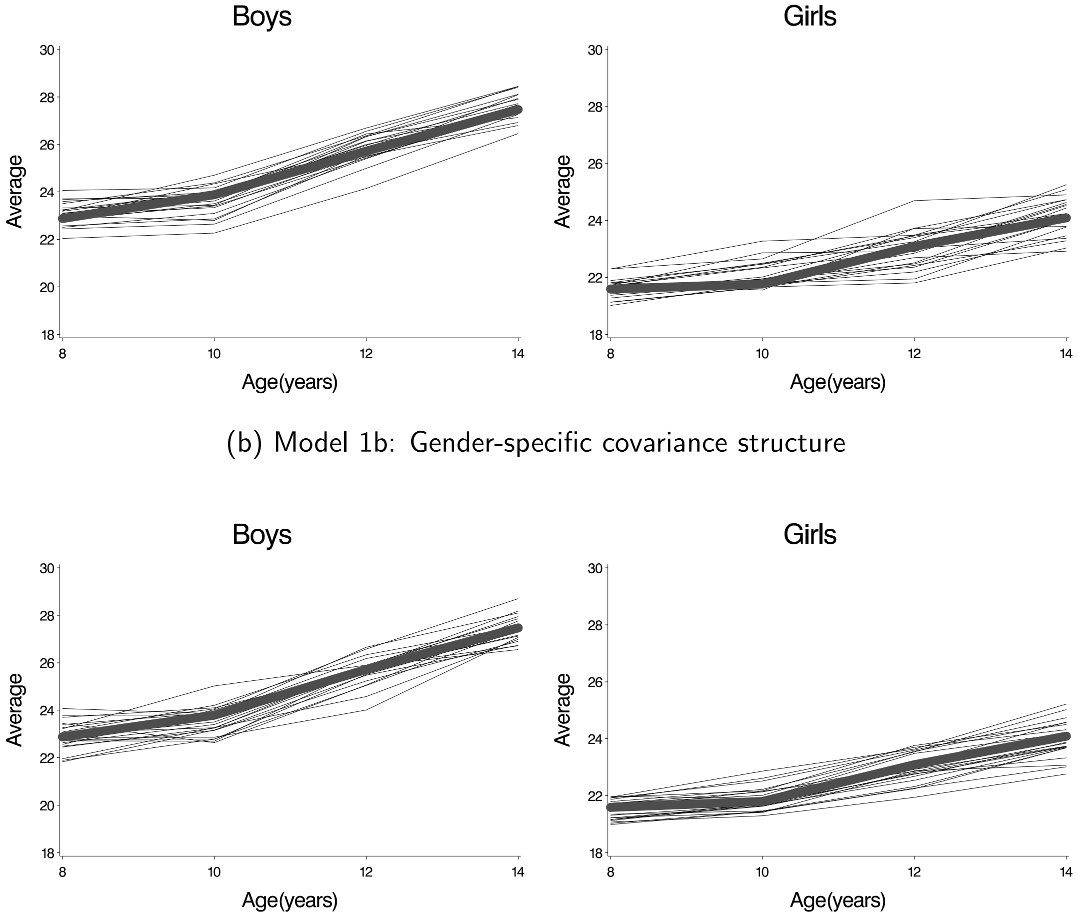
(c)

| | | **model** | |
|---|---|---|---|
| | | complete | incomplete |
| **raw** | complete | | |
| **data** | incomplete | $*$ | |

(d)

| | | **model** | |
|---|---|---|---|
| | | complete | incomplete |
| **raw** | complete | | |
| **data** | incomplete | | $*$ |

Figure 6: *Model assessment when data are incomplete. (a) Two dimensions in model (assessment) exercise when data are incomplete. (b) Ideal situation. (c) Dangerous situation, bound to happen in practice. (d) Comparison of data and model at coarsened, observable level.*


(a) Model 1a: Equal covariance structure

(b) Model 1b: Gender-specific covariance structure

Figure 7: *The Orthodontic Growth Data. Sample averages for the augmented data (bold line type), compared to sample averages from 20 simulated datasets, based on the method of Gelman* et al *(2005). Both models assume a saturated mean structure and compound symmetric covariance. Model 1a assumes the same covariance structure for boys and girls, while Model 1b allows gender-specific covariances.*

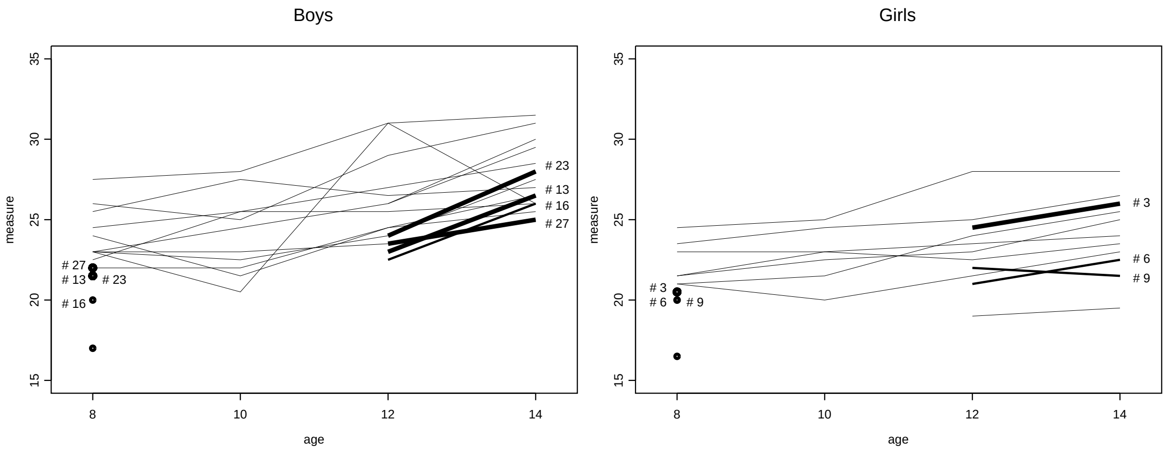


Figure 8: *The Orthodontic Growth Data. Individual profiles of the incomplete version of the data, with highly and moderately influential subjects highlighted by more and less boldface line type, respectively.*

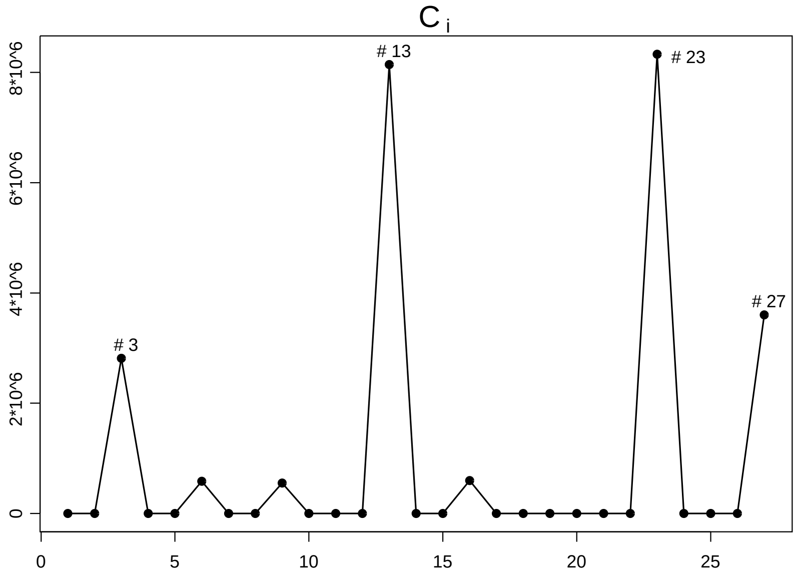


Figure 9: *The Orthodontic Growth Data. Local influence measures.*

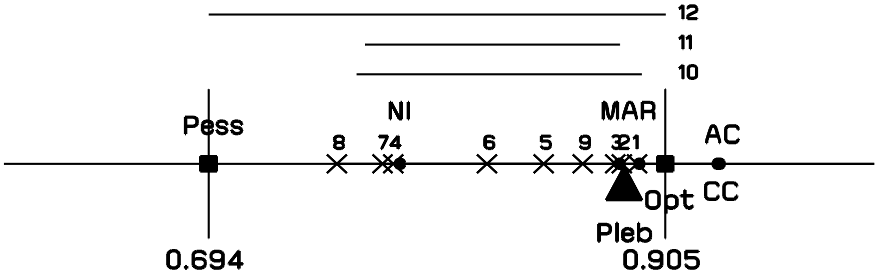


Figure 10: *The Slovenian Public Opinion Survey. Relative position for the estimates of "proportion of YES votes", based on the models considered in Rubin, Stern, and Vehovar (1995), and on the BRD Models. The vertical lines indicate the nonparametric pessimistic-optimistic Bounds. (Pess: pessimistic boundary; Opt: optimistic boundary; MAR: Rubin et al. 's MAR model; NI: Rubin et al. 's MNAR model; AC: available cases; CC: complete cases; Pleb: plebiscite outcome. Numbers refer to the BRD models. Intervals of ignorance (Models 10–12) are represented by horizontal bars.)*