23

REPEATED MEASURES DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Recognise and understand the unique characteristics of a repeated measures structure.
2. Use descriptive and graphical tools to quantify and visualise the repeated measures structure of a dataset.
3. Use simple univariate approaches to analyse repeated measures data.
4. Use mixed models to analyse repeated measures data, and understand the limitations of random-intercept mixed models for such data.
5. Choose among a variety of correlation structures that might be appropriate for repeated measures or spatial data.
6. Understand the fundamental differences between mixed model and generalised estimating equation (GEE) approaches for analysis of clustered data.
7. Use GEE procedures to analyse clustered data, in particular repeated measures data.

23.1 INTRODUCTION

In this chapter, we describe methods for analysis of repeated measures data that, as discussed in Chapter 20, could be considered as a special type of clustered data. It is also one of the most commonly encountered data structures in epidemiology and the health sciences in general. A wide selection of methods and approaches exist for analysis of such data, and the choice between them depends on the characteristics of the data at hand as well as the objective of the analysis. We cannot, within a single chapter in the book, cover all methods, or cover the methods selected in full detail. Among the many excellent textbooks on repeated measures or longitudinal data, a standard (fairly theoretical) reference is Diggle *et al* (2002), and also Fitzmaurice, 2004; Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2001 provide extensive coverage in a blend of theory and practice.

To illustrate the methods, we will revisit the blood pressure data (-bp-) studied in Chapter 21 and explore analyses involving all patient visits post treatment (up to four visits, denoted 3–6) instead of the single visit considered in Chapter 21.

23.1.1 What are repeated measures data?

A **longitudinal** study can be characterised by having several measurements over time on the same **subjects** (individuals, or sometimes other experimental units such as sample plots in a field), as opposed to studies with only one measurement per subject. A longitudinal study certainly involves **repeated measures** (or measurements) on the same subjects, but the latter term is sometimes used in a slightly more general sense to denote consecutive measurements, *ie* measurements with a certain inherent ordering not related to time (*eg* measurements obtained across a range of doses or magnifications). If there is no ordering to multiple measures on the same subjects, we might think of these as clustered within the same subject instead, as discussed previously (Chapters 20–22).

From the clustering of measurements within subjects, we already know that it is usually unreasonable to consider the measurements as independent; by doing so we would ignore any subject characteristics affecting our outcome. For example, in growth curves (one particular example of repeated measures) a subject that is large relative to its fellow subjects at a young age would tend to remain relatively large throughout the growth period. The (time) ordering of the measurements introduces another type of association between measurements because usually 2 measures on the same subject that are taken close together in time will be more strongly associated than measures taken further apart in time. Again in growth curves, the initial weight (say) should be more strongly correlated with the immediately ensuing growth measurements than the last ones. Such a pattern of correlations we broadly refer to as **autocorrelation** without stating specifically how the correlation is reduced by increasing time distance. It is because of this feature of repeated measures data that they cannot generally be treated as a hierarchical data structure. Specifically, a 2-level hierarchical structure (with measurements nested within subjects) does not take time ordering of the measurements into account in the random part of the model. Where individuals within a group (*eg* family members) can be interchanged without altering the meaning of the data, observations over time on the same subject cannot. Despite the intuitive logic of autocorrelation, some outcomes may not show any autocorrelation, so we will need to assess for each dataset individually whether autocorrelation or a simple clustering within subjects is present in the data, or perhaps no clustering at all.

As the range of methods that can be applied to a dataset depends on its structure, it is useful to introduce some terminology to describe repeated measures data. The most regular data type has the same number of measures taken for each subject (*ie* is **balanced** over time) with **uniform** (*ie* the time points are the same across subjects) and equally spaced (**equidistant**) time points. Protocols in clinical trials commonly require equidistant sampling or follow-up. The presence of **missing data** will make designs unbalanced, but are difficult to avoid. A unique feature to repeated measures data is that observations may be missing because the subject exits prematurely from the study (**drop-outs**). For example, the blood pressure trial was planned to be balanced with uniform but non-equidistant time points, because visits 3–5 were two weeks apart, whereas visit 6 was 6 weeks after week 5. In practice, some patients left the study before completing all visits (drop-outs), and the visits may not have taken place exactly at the planned intervals. Generally speaking, the most regular data types will not only allow a wider range of analytical approaches, but will also be easier to analyse.

23.1.2 Descriptive statistics and graphical displays

As the choice of analytical methodology for repeated measures data will also depend, to some extent, on the characteristics of the data at hand, it is crucially important to familiarise oneself with the data before plunging into a complicated analysis. Two obvious approaches for that are suitably chosen descriptive statistics and visualisations of the data, and analysis by simple (possibly simplistic) procedures such as those described in Section 23.2. To begin with, one should assess the distribution of time points within each subject to determine how regular these are (*eg* balanced, equidistant). Next, one should compute suitable means across subjects at different time points to get an impression of how time affects the outcome; these can be plotted against time in a **mean plot**. If time points are uniform, it is often also useful to compute the crude correlations between measurements at different time points. This will often require you to shift from **long data format** (each row corresponds to one measurement) to **wide data format** (each row corresponds to one subject, with the measurements distributed across several columns). Finally, it is recommended to construct one or multiple **profile plots** showing the series of observations over time on the subjects. If there are too many subjects to display them all in a single plot, one may construct plots for suitably chosen groups (formed by predictor values) and/or select some subjects for display. Examples 23.1–2 illustrates the approach for the blood pressure dataset; Example 23.1 is restricted to patients with complete records and Example 23.2 specifically focuses on incomplete series of repeated measures.

23.1.3 Longitudinal versus cross-sectional study designs

Diggle *et al* (2002), Chapter 2 contrasted the longitudinal and cross-sectional designs and we'll briefly review the main points. A cross-sectional study can be used to give information about differences between subjects in different subpopulations; in addition, a longitudinal study can give information about changes in subjects over time. This is particularly important if we want to assess the impact of predictors that change over time. In a cross-sectional study, these can only be estimated from between-subject regressions, and in order to interpret them as changes within an individual, we would need to assume that the within-subject regression has the same slope, *ie* that the predictor has no contextual effect (Section 21.4).

In addition, longitudinal designs can be substantially more powerful statistically than cross-sectional designs for inference about within-subject predictors. This is analogous to the gain of

a block design with treatments allocated within blocks, *eg* a cross-over design. For between-subject predictors, the cross-sectional design with its between-subject independence is the most powerful if the cost of sampling different subjects is not larger than the cost of repeatedly sampling the same subject, by the same reasoning as in Section 20.3.3.

23.2 Univariate and multivariate approaches to repeated measures data

In this section, we will briefly review some relatively simple statistical procedures to deal with repeated measures data in regular between-subject designs with independent subjects. That is, we assume balanced and uniform series on all subjects, and consider inference about predictors

Example 23.1 Graphs and descriptive statistics for blood pressure data
data = bpwide

Initial explorations of the data will suggest a suitable scale for analysis. In Chapter 21, we established there was no need for transformation of the outcome, diastolic blood pressure (dbp) (at visit 3), so we present graphs and descriptive statistics on original scale. For this clinical trial to compare 3 treatments, it is natural to compute descriptive statistics separately per treatment group and visit, including the initial visit prior to treatment. (Visit 2 was the visit at which the treatment was initiated.) The results are restricted to 256 (88.9%) patients with complete records.

Mean dbp (SD)	Number of	Visit				
treatment	patients	1	3	4	5	6
Carvedilol	82	102.9 (4.8)	94.8 (8.4)	92.3 (8.4)	92.4 (8.8)	90.4 (8.8)
Nifedipine	84	102.1 (4.3)	92.8 (9.3)	91.9 (8.6)	89.5 (8.8)	90.1 (7.8)
Atenolol	90	103.0 (4.8)	91.4 (9.3)	91.1 (9.2)	88.4 (9.7)	88.6 (8.7)

The table shows a large drop in blood pressure after treatment onset in all treatment groups. Differences between treatments groups are small but apparently consistent over time. The low SD at visit 1 is probably caused by the selection criteria for patients to be enrolled in the study. Fig. 23.1 shows a profile plot for 15 selected patients and a mean plot comparing treatment groups.

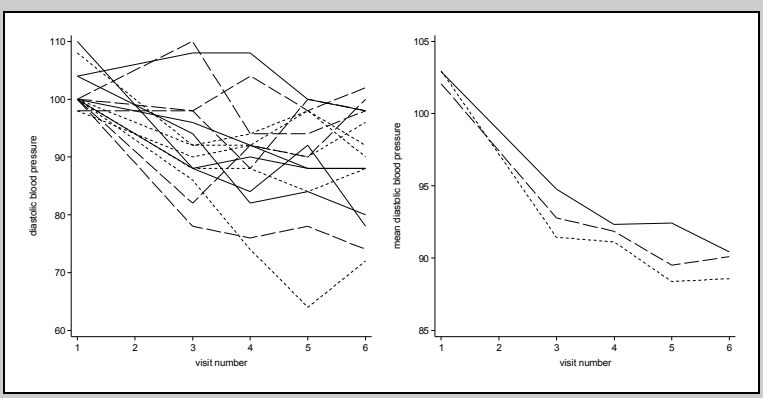

Fig. 23.1 Profile plot (left) and mean plot (right) for patients in 3 treatment groups: Carvedilol (solid), Nifedipine (long dashes) and Atenolol (short dashes)

(*continued on next page*)

Example 23.1 (*continued*)

The profile plot shows quite noisy curves with a downward trend, and no clear evidence of high within-subject correlation which would manifest itself by subjects tending to remain high or low throughout the time period (this visual phenomenon is sometimes called 'tracking'). The mean plot shows higher blood pressures for the new drug, Carvedilol, than the two other drugs, (even prior to treatment (visit 1) compared to Nifedipine), although the difference is smallest at visit 1. We also present in tabular form below the simple correlations (left), variances and covariances (right) of the 5 measures on the same patients (see Sections 20.1 and 23.2.3).

	Correlations					Variances/Covariances				
Visit	**1**	**3**	**4**	**5**	**6**	**1**	**3**	**4**	**5**	**6**
1	1					21.62				
3	0.311	1				13.16	82.61			
4	0.333	0.570	1			13.51	45.28	76.29		
5	0.245	0.511	0.624	1		10.52	42.96	50.39	85.41	
6	0.149	0.477	0.518	0.607	1	5.86	36.75	38.34	47.54	71.86

For example, the correlation between measurements at the first and last visits is 0.149, and the variance of measurements at the first visit is 21.6 and, as we already noted, larger than at later visits. The correlations with the initial dbp measurements are smaller than among the follow-up measurements, reflecting variation in patients' responses to treatment. As expected, the data show autocorrelation because correlations drop down as time between visits increase, however only moderately so.

at the subject level; the blood pressure trial data is (apart from the missing values) an example of such a data structure if we allow ourselves to disregard any impact of the centres. These methods are less commonly used than those in the following sections, in part because of their demands on the data structure and because they may not fully use the information in the data. Nevertheless, they may serve as reference points for more complicated analyses, and could in some situations suffice to draw conclusions about the study hypotheses.

23.2.1 Univariate methods

We call these methods 'univariate' because they essentially avoid modelling the repeated measures structure by reducing the within-subject series of measurements to one (or several) statistics computed for each subject. With one observation (the computed statistic) per subject, a multivariable analysis involving any predictors and further hierarchical structure for the subjects follows the lines of previous chapters of the book.

The most basic procedure is analysis by **separate time points**. In the blood pressure trial, there might be particular interest in comparing the treatments at the first and last post-treatment visit (visits 3 and 6). With the single categorical predictor treatment, the analysis would be a linear model corresponding to a 1-way ANOVA (Chapter 14). The analysis would not be wrong, but it would be inefficient because all the preceding measurements are not used. If a similar analysis was carried out for each of the time points, it might become difficult to combine the conclusions from the different analyses. As the data from different time points were treated separately, we would not know how strongly correlated they were, and then we could not tell whether a few significances at different time points strengthen the evidence of predictor effects,

Example 23.2 Incomplete series of records in blood pressure data
data = bp

The patterns of missing outcomes in repeated measures data is often informative for assessing whether missing data, and in particular drop-outs, could be associated with characteristics of either the outcome or predictors (which would create a risk for biased results, see Section 15.5). Among the 288 patients in the blood pressure trial, 32 patients had incomplete records, and their observation patterns are cross-tabulated by treatment groups. The table gives counts of patients with different series of incomplete records across visits 3–6; for example, 8 patients in the Carvedilol group had records at visits 3 and 4 only.

Visit				Treatment group			
3	4	5	6	Carvedilol	Nifedipine	Atenolol	Total
0	1	1	1	0	0	1	1
1	0	0	0	5	2	1	8
1	0	1	1	1	0	0	1
1	1	0	0	8	3	1	12
1	1	1	0	4	4	2	10
1	1	1	1	82	84	90	256
	Total			100	93	95	288

The table shows that most (30) of the incomplete series are due to drop-outs, which in turn are not quite evenly distributed between treatment groups: most occur in the Carvedilol group. The profile plot for 20 of the incomplete series in Fig. 23.2 shows similar trend and variability as the profile plot for complete series in Fig 23.1. The main concern of the missing data seems therefore to be the overrepresentation by the Carvedilol group. This feature of the data should be investigated independently of the actual recorded data.

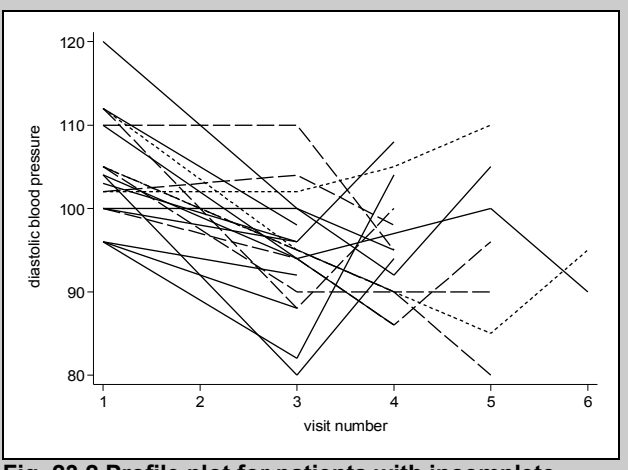

Fig. 23.2 Profile plot for patients with incomplete series of records in 3 treatment groups: Carvedilol (solid), Nifedipine (long dashes) and Atenolol (short dashes).

or whether we essentially just saw the same evidence several times. On the other hand, running the analysis at multiple time points could increase our Type II error, in particular if we fell victim to the temptation of selecting the time points with the 'best' effects. This discussion shows the need for a formal rule for managing any selection of time points for analysis; one general applicable rule is the **Bonferroni correction** for performing multiple analyses (Section 14.10.1) by dividing the significance level by the number of analyses performed or the number of time points considered for selection for analysis. This approach produces a valid analysis for separate time points but it is a weak analysis, in part due to the (conservative) Bonferroni correction. More importantly, the analysis does not use the longitudinal information in the data (the subjects at different time points could be different), and it does not describe or analyse the development over time. Example 23.3 demonstrates the approach applied to the blood pressure data.

Analysis by a **summary statistic** (also, response feature or derived variable) is a refinement of the time point method, performed in 2 steps. In the first step, you choose a single quantity to calculate from each subject's profile, for example the drop in blood pressure from the first to last measurement. This again results in a single observation per subject on which we then carry out a between-subject analysis as above. The effectiveness of the approach depends on whether one can devise a good summary statistic that captures the relevant information inherent in the profiles; the choice is usually guided by inspection of profile plots. Some standard choices of summary statistics are: the subject mean or median, the within-subject slope, the gain (or drop),

Example 23.3 Univariate methods for blood pressure data
data = bpwide

The subset of patients with complete records is used to illustrate univariate methods, while, for the purpose of illustration, disregarding the centres. The measurements at each time point were analysed by a linear model to compare the treatment groups while, for visits 3–6, adjusting for the initial blood pressure (in order to reduce between-subject variation). Analysis of pre-treatment outcomes is mostly of interest to ascertain that the randomization created comparable groups. Therefore, the Bonferroni adjustment should involve the 4 subsequent time points; thus, P-values less than 0.05/4=0.0125 could be considered as significant. In addition, 2 summary statistics were explored: the immediate drop (difference between blood pressures at visits 1 and 3) and the slope for visits 3–6 expressing trends of longer term treatment benefit, and also motivated by the roughly linear profiles in Fig. 23.1. The table gives P-values for the effects of treatment and the initial blood pressure.

P-values		Separate analysis at visit					Summary statistic	
Effect	df	1	3	4	5	6	drop	slope 3–6
treatment	2	0.377	0.038	0.501	0.011	0.252	0.036	0.917
initial dbp	1	n/a	<0.001	<0.001	0.001	0.014	n/a	0.004

The analyses at separate time points showed significant treatment effect only at visit 5 (after Bonferroni correction). The group means were given in Example 23.1. As already noted there, the means for the Carvedilol group are higher than for the other treatment groups, with the differences being largest at visit 5. The regression coefficients for the initial dbp were all positive, but decreasing in magnitude at later visits, as one would expect. The immediate drop in blood pressure after treatment shows a significant difference between treatments; if the initial blood pressure was included as a covariate in this analysis, the inference for treatments would be equivalent to that for visit 3. Finally, there is absolutely no indication that the longer term trend differs between treatments.

and the area under the curve (AUC). Summary statistics should generally be chosen to have interpretations of practical and/or scientific interest. They are not (primarily) based on statistical considerations; for example, the use of the within-subject slope from a regression does not require that the curve is modelled well by a straight line (or a statistical assessment of linearity). The slope can simply be used as a measure of average increase even if there is some curvature. For growth curves, average gain (per time unit) is a standard growth measure, even if the growth may show some non-linearity. We also illustrate the use of summary statistics in Example 23.3.

Advantages of the summary statistic approach are its simplicity and flexibility, including its potential use for discrete data, and the direct access to features of interest that may be difficult to extract from complex models. Suitably chosen summary statistics can be both powerful and robust towards model assumptions and data irregularities (Everitt, 1995; Senn *et al*, 2000). Disadvantages are the subjective choice of the statistic, the loss of information by reducing each profile to a single statistic, and the limited information provided by the analysis (*eg* no correlations or predictions). Also, it is difficult to incorporate strong or key within-subject predictors into the approach.

23.2.2 Repeated measures ANOVA

Treating the repeated measures within subjects as a hierarchical structure leads to models with subject random effects. The simplest of such models, the random-intercept model, can in regular between-subjects designs be analysed with the ANOVA-based approach for mixed models (Section 21.5.1), and is sometimes termed the 'split-plot' approach to repeated measures data, referring to the link between hierarchically structured data and a split-plot design explained in Section 20.2.1. We saw in Chapter 21 that, in a random-intercept model, the correlations are the same (and positive) between all units within a cluster (*eg* Eq 21.4); in Section 23.3.1, we will call this a **compound symmetry** correlation structure. But we noted already in the introduction that we would expect autocorrelation to be present in repeated measures data, so a random-intercept model induces the wrong correlation structure. In essence, the hierarchical model fails because it does not take into account the time ordering of the repeated measures on each subject. For these reasons, the random-intercept model is by now considered inadequate for most repeated measures data analyses. It is, however, perfectly valid for data with only 2 repeated measures per subject, and may give a reasonable analysis for short series (with 3 or 4 time points) because compound symmetry may not be far off in such cases. As the random-intercept model provides the simplest analysis of the full dataset, it is often used as a starting point (or reference) for further more complex models. It could also be used for first decisions about the modelling that are unlikely to require an accurate correlation structure, *eg* a choice of transformation of the outcome (although one would be advised to reassess the transformation with the final model).

As the first of several approaches to assess (test) the assumed correlation structure, we describe the **repeated measures ANOVA** method for regular between-subjects designs. The aim of this method is to assess, and possibly adjust, the impact of the assumed correlation structure on the test statistics of the ANOVA table; thus, it is essentially a method for correcting test statistics but does little to adjust other features of the statistical inference such as standard errors on estimates. For this reason, and because of the design requirements (which *eg* imply that missing values cannot be managed in any easy way), the repeated measures ANOVA approach has largely been superseded by extensions of the mixed modelling approach (Section 23.3). We mention it here mainly because of the insights it offers into the impact of wrongly assuming a

compound symmetry structure. It can be shown (in regular designs) that a violation of the compound symmetry assumption affects only within-subjects effects (*ie* effects involving time) and makes the corresponding uncorrected test statistics of the ANOVA table too liberal (*ie* gives a P-value that is too small). Several correction factors exist to adjust (reduce) the degrees of freedom of the *F*-statistics to achieve approximately correct inference. We illustrate the procedure in Example 23.4. As already mentioned, the general consensus among statisticians is that the repeated-measures ANOVA approach offers no real advantages in return for its strong restrictions, and we therefore do not recommend it for general use.

23.2.3 Multivariate analysis

Multivariate statistical methods apply to data where the measurement on each subject, or experimental unit, consist of multiple records instead of a single record, as has been the case in previous chapters of the book. As an example, patients often undergo a panel of tests at check-ups. If such multiple records are compiled into a vector of observations on each subject, we can think of **multivariate data** as consisting of vectors of observations instead of single observations per measurement. A large body of theory and methods exist for multivariate data, but we concentrate on how they can be applied to repeated measures data (Davis, 2002), Chapters 3–4). Repeated measures on the same subject may be considered as a single observation (vector), consisting of the entire set of values across time points. Let us introduce a bit of notation to support the idea: Y_{ij}=measurement for subject i at time j, where there are m time points j=1,...,m. Then, in a multivariate framework, the basic observation for subject i is the vector Y_i=(Y_{i1},...,Y_{im}). Multivariate linear models extend the usual linear models (Chapter 14) by modelling the observation vector in terms of its mean (vector) and variance (matrix). The mean vector consists of the mean outcome at different time points, and the variance matrix consists of the variances at and the covariances between the different time points. It is more common to refer to the latter matrix as the **covariance matrix** (sometimes also variance-covariance matrix), so we'll use that term. Also, the covariances are more intuitive to interpret when rescaled as correlations (for the relation between covariance and correlation, see Eq 20.1). For a set of measurements (Y_1,...,Y_m) on the same subject (where, for simplicity, we suppress the subject indicator i), the covariance matrix cov(Y) and the correlation matrix corr(Y) are the (mxm)-matrices holding all the covariances, or correlations, between pairs of measurements:

$$\mathrm{cov}(Y)=\begin{pmatrix} \mathrm{var}(Y_1) & & & & \\ \mathrm{cov}(Y_1,Y_2) & \mathrm{var}(Y_2) & & & \\ \mathrm{cov}(Y_1,Y_3) & \mathrm{cov}(Y_2,Y_3) & \mathrm{var}(Y_3) & & \\ \vdots & \vdots & \vdots & & \vdots \\ \mathrm{cov}(Y_1,Y_m) & \mathrm{cov}(Y_2,Y_m) & \mathrm{cov}(Y_3,Y_m) & \cdots & \mathrm{var}(Y_m) \end{pmatrix} \qquad \textit{Eq 23.1}$$

$$\mathrm{corr}(Y)=\begin{pmatrix} 1 & & & & \\ \mathrm{corr}(Y_1,Y_2) & 1 & & & \\ \mathrm{corr}(Y_1,Y_3) & \mathrm{corr}(Y_2,Y_3) & 1 & & \\ \vdots & \vdots & \vdots & & \vdots \\ \mathrm{corr}(Y_1,Y_m) & \mathrm{corr}(Y_2,Y_m) & \mathrm{corr}(Y_3,Y_m) & \cdots & 1 \end{pmatrix} \qquad \textit{Eq 23.2}$$

The matrices are symmetric, so for clarity, the values above the diagonal have been left blank. In Example 23.1, we displayed the covariance and correlation matrices for the 5 repeated measures of diastolic blood pressure per patient in the clinical trial.

Multivariate analysis of variance (MANOVA) assumes normally distributed multivariate outcomes with a mean modelled in terms of subject-level predictors and a covariance without any specific structure (although covariances may be assumed to be either constant or heterogeneous across predictor groups). Thus, the model defaults to heterogeneous variances across time points and makes no assumptions about the correlation structure. The analysis provides estimates of means (with SEs), variances, and correlations, and these can be used to test hypotheses about the effects of between-subject predictors and time. Different test statistics exist for the same hypotheses (although for simple hypotheses they'll coincide); Wilk's lambda is a sensible overall choice, with reference F-distributions. The standard test statistics offered by MANOVA software do not include hypotheses related to time (because in general multivariate analysis, there is no structure among the multivariate responses), so these may need to be set up manually by specifying suitable contrasts; for details consult Davis (2002; Chapter 3–4). The multivariate analysis is similar to analysis by a mixed model with unstructured correlations (Section 23.3.2), which (with suitable statistical software) gives easier access to specific contrasts and tests. Example 23.4 gives results from both the multivariate and mixed-model analysis for the blood pressure data.

One advantage of the multivariate approach is that it avoids any problems with a wrongly specified correlation structure (as in the random-intercept model). However, this advantage also contains the potential drawback that estimation of many covariance parameters (all the variances and correlations) may be ineffective or outright impossible, especially with long series of measurements. Davis (2002), Chapter 6 cites simulation studies with small/moderate number of subjects that have shown the multivariate approach to provide exact and better statistical inference than analysis by the mixed model. The main drawbacks of the multivariate approach as presented here are its strong requirements: normally distributed, balanced data with uniform time points, no missing values and no within-subject predictors. However, one software implementation (MLwiN) relaxes all these conditions and also allows for additional hierarchical structure (Rasbash *et al*, 2008, Chapter 14).

23.3 Linear mixed models with correlation structure

Having noted in Section 23.2.2 the deficit of the simplest linear mixed model, the random-intercept model, for repeated measures data, we will discuss here 2 ways of extending the model to incorporate more realistic correlations for continuous repeated measures data. With both of these extensions (in Sections 23.3.2–3) the model would still be termed a linear mixed model, so the important details lie in the actual specification of the model. Also, the advantages of the linear mixed model, such as its flexibility to handle hierarchical structure and predictors at multiple levels as well as its likelihood-based inference and resulting robustness to missing values—as long as these are missing at random (Section 15.5)—will remain intact. These are some of the distinct advantages of the linear mixed model approach over the simpler approaches reviewed so far. In this section, we will not revisit the entire analysis of the linear mixed model from Chapter 21 but concentrate on describing the 2 extensions of the model and their impact on the analysis.

Example 23.4 ANOVA and MANOVA methods for blood pressure data
data = bpwide

We continue the illustration of methods by the blood pressure data from Examples 23.1 and 23.3, still, for the purpose of illustration, disregarding the centres. Here we show results from a random-intercept model, its repeated measures ANOVA adjustment, a multivariate (MANOVA) analysis and a mixed model with unstructured covariance matrix (Section 23.3.2). Due to the balancedness of the complete subset of the data selected, the parameter estimates for all models are the means for the treatment by visit combinations shown in Example 23.1. The table below gives P-values for fixed effects hypotheses for the categorical predictors in 2 versions of the data, including either all visits or only the post-treatment visits (3–6). The assumed variance homoscedasticity of the random intercept and repeated measures ANOVA analyses is clearly violated when all visits are included. For visits 3–6, the estimated within- and between-subject variances in the random-intercept model were 35.41 and 42.83, respectively, corresponding to an *ICC* of 0.55. The repeated measures ANOVA gave an estimated Huyhn-Feldt correction factor of ε=0.98 (where ε=1 means that no adjustment to test statistics is required because of violations of the assumed correlation structure of the random-intercept model), and adjusted the *F*-distribution degrees of freedom by multiplication with this factor; for example, the adjusted degrees of freedom for tx*visit were (ε*6, ε*759)=(5.9, 744). With ε so close to 1, the adjustment is minimal.

P-values		Model / Method			
Effect	df	random intercept	rep. meas. ANOVA	MANOVA	mixed, unstruct.
Visits 1–6					
tx	2	0.084	0.084	0.070*	0.084
visit	4	<0.001	<0.001	<0.001	<0.001
tx*visit	8	0.060	0.060	0.045	0.043
Visits 3–6					
tx	2	0.061	0.061	0.068*	0.061
visit	3	<0.001	<0.001	<0.001	<0.001
tx*visit	6	0.245	0.246	0.124	0.119

*simultaneous test across all time points

The estimated error covariance matrix for the MANOVA and the mixed model were identical and very close to the values shown in Example 23.1 (not shown). Contrasting the correlations with the single estimated correlation (*ICC*) of the random-intercept model illustrates that the random-intercept model is appreciably off the actual correlation structure. However, the impact on test statistics was limited (though as some P-values traversed the significance level of 0.05, the impact of conclusions could be substantial if the P-values were interpreted too rigidly). Note that the first MANOVA test (in both settings) is for a different hypothesis (involving all time points simultaneously) and cannot be compared with those of the other methods.

The conclusion about the comparison of treatments is that during visits 3–6 the treatment effects do not change strongly, and overall the difference between treatments is close to significant. It is not surprising that the tx*visit interaction is stronger (and significant) when the pre-treatment visit is included where there should be no treatment effects (and Example 23.3 indicated this to be the case). One substantial weakness of the present analyses for visits 3–6 is that the pre-treatment dbp was not included; this weakness pertains in fact only to the repeated measures ANOVA and MANOVA methods for, as Example 23.5 shows, the mixed models do allow for inclusion of this predictor. We could use multiple comparisons to explore the significant tx*visit effect in the first setting, but if the significance is mostly due to the inclusion of pre-treatment comparisons, it seems more natural to pursue analysis for visits 3–6 only.

23.3.1 Correlation structure

Before we proceed to the model extensions, we will distinguish more precisely than we have done so far between the different correlation structures of importance for the modelling. First, despite our usage of the term 'correlation structure', we really mean the structure of the covariance matrix, because the variances are equally part of the structure we're modelling (however, 'covariance structure' seems less intuitive than correlation structure). Second, the correlation structure targeted by the modelling, and which violation may affect the inference, is that of the errors, not the structure of the observed data. The difference between the 2 is that, in the former, the fixed and random effects of the model have been estimated and adjusted for (*ie* eliminated from the correlations). We discuss the impact of fixed and random effects in turn.

If the fixed effects are strong, the crude and adjusted correlations can be appreciably different. We noted in Example 23.4 that the crude correlations were very similar to those estimated for the errors of the model; in these data, the fixed effects were indeed fairly small (within each time point). In order for fixed-effects predictors to account for some of the anticipated autocorrelation in repeated measures data, they must include a within-subject predictor that itself shows autocorrelation. One possible example, for studies involving a substantial time span, is the subject's age, but one should be aware of the potential collinearity with study time upon inclusion of subject intercepts. Realistically, though, it is not common that the fixed effects eliminate substantial parts of the autocorrelation.

Random effects of random-intercept type can, as we have seen, only induce compound symmetry correlations. Random slopes however can induce autocorrelation if the predictor involved is correlated with time. One obvious candidate for a random slope with potential to induce autocorrelation is therefore ... time. Adding a random slope with time to a random-intercept model induces autocorrelation, and may therefore remove autocorrelation from the errors. Linear mixed models with random slopes for time are also called **trend** models, and constitute one of the 2 extensions of the random-intercept model to deal with autocorrelation (Section 23.3).

We will next describe a range of correlation structures that can exist either for repeated measures data or for the errors in a model of such data. Table 23.1 lists some of the more common correlation structures for repeated measures in the case of m=4 repeated measures on the same subject. For simplicity, we show only the correlation matrix in all cases except the last one. However, if variances are assumed to be equal (σ^2), the covariances are simply the correlations multiplied by σ^2.

The first 2 correlation structures are well-known and included mainly to familiarise the reader with the display. Recall that the correlation ρ in the **compound symmetry** structure induced by a random-intercept model can be expressed in terms of the variance components σ_g^2 and σ^2 as $\rho=\sigma_g^2/(\sigma_g^2+\sigma^2)$ (Eq 21.4). The alternative name, an **exchangeable** structure, refers to the fact that since correlations are the same all over, the units (here the time points, but in our hierarchical models the individuals within a cluster) can be interchanged (or exchanged) without affecting the structure.

Table 23.1 Repeated measures correlation structures for four repeated measures/subject

Name	Correlation structure	Interpretation
uncorrelated or independent	$\text{corr}(Y)=\begin{pmatrix}1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1\end{pmatrix}$	uncorrelated (for normal data: independent) observations
compound symmetry, or exchangeable	$\text{corr}(Y)=\begin{pmatrix}1 & & & \\ \rho & 1 & & \\ \rho & \rho & 1 & \\ \rho & \rho & \rho & 1\end{pmatrix}$	hierarchical, mixed model (same correlation between all pairs of observations)
ar(1), or first order autoregressive	$\text{corr}(Y)=\begin{pmatrix}1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1\end{pmatrix}$	repeated measures or time-series model with power decay of correlations
arma(1,1), or first order autoregressive moving average	$\text{corr}(Y)=\begin{pmatrix}1 & & & \\ \gamma & 1 & & \\ \gamma\rho & \gamma & 1 & \\ \gamma\rho^2 & \gamma\rho & \gamma & 1\end{pmatrix}$	extended repeated measures or time series model with power decay
Toeplitz, or stationary	$\text{corr}(Y)=\begin{pmatrix}1 & & & \\ \rho_1 & 1 & & \\ \rho_2 & \rho_1 & 1 & \\ \rho_3 & \rho_2 & \rho_1 & 1\end{pmatrix}$	repeated measures with unconstrained correlations at different spacings
unstructured	$\text{corr}(Y)=\begin{pmatrix}1 & & & \\ \rho_{12} & 1 & & \\ \rho_{13} & \rho_{23} & 1 & \\ \rho_{14} & \rho_{24} & \rho_{34} & 1\end{pmatrix}$	repeated measures with entirely unconstrained correlations
unstructured with inhomogeneous variances, or non-stationary	$\text{cov}(Y)=\begin{pmatrix}\sigma_1^2 & & & \\ \sigma_{12} & \sigma_2^2 & & \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2\end{pmatrix}$	repeated measures, unconstrained variances and correlations

The simplest structure showing the desired decay in correlation with increasing distance between observation is first order **autoregressive**, or **ar(1)**, in terminology originating from time series analysis (Section 14.11). It involves 2 assumptions: that all pairwise correlations that are a certain number of time steps (or points) apart are correlated to the same degree, and that the correlations decay as powers of the number of time steps that separate 2 observations. The first assumption implies for example that the correlation in the observation pairs (1,2), (2,3) and (3,4) are all the same (and equal to ρ). The assumption is sometimes called **homogeneous** or stationary correlations (discussed further below). **Note** The decay of ar(1) correlations is quite rapid; *eg* for ρ=0.5, observations 4 time steps are close to uncorrelated (0.5^4=0.0625).

More complex correlation structures than ar(1) are often useful as well, in order to incorporate

either a slower or less consistent decay of correlations with time distance. The **arma(1,1)** structure, also originating from time series analysis, and the **Toeplitz** structures accomplish this with 1 and (m-2) additional parameters, respectively. The choice (and nomenclature) of correlation structures available for modelling depends on the statistical software; many additional and more complex structures exist, but the 3 homogeneous structures mentioned here are usually always available. **Homogeneous** structures are most meaningful if the time points are equidistant. In some situations, when processes occur at different speeds in different stages of the time period considered, one could perhaps argue non-equidistant time points to be 'biologically equidistant' (*ie* that they should have the same impact). For example, if one studies the impact of the injection of a pharmaceutical into animals (or humans), one may choose to measure the response at follow-up times 1, 2, 5, 10, and 30 minutes post-injection. If the biological processes happen much more quickly in the initial phase after injection, it may still be meaningful to assume homogeneous correlations. Clearly such reasoning would require strong biological justification.

An **unstructured correlation structure** will let the data speak for itself; we already saw its use in the blood pressure data (Example 23.4). The drawback of this structure is that, with a long series of repeated measures, the number of parameters involved grows so large that they become difficult to estimate and interpret. Another question is whether heterogeneous variances across the time points should be assumed. All correlation structures have a corresponding version with heterogeneous variances (but it may not be implemented in the software). There are often good biological reasons why variances should not be assumed constant over time; on the other hand, heterogeneous variance structures will also increase the number of parameters appreciably for a long series.

One note of caution about correlation structures: you need to ensure that the time points are properly understood by your software, in particular if the data contain incomplete series (due to missing values). For example, it makes a difference with most correlation structures whether recordings were taken at times (1,2,3,4), at times (3,4,5,6) or at times (1,2,5,6). If the time points are not uniform across subjects (with allowance for missing values), at least to a good approximation, the correlation structures will not be meaningful across the dataset, and modelling based on certain fixed correlation structures will be misleading. Such data structures therefore raise the need to incorporate into the matrices the actual recording times.

For non-equidistant repeated measures or spatial data, denote by $d_{jj'}$ the distance between observations j and j'. For longitudinal data where locations correspond to time points, the $d_{jj'}$ would be the (absolute) difference between the recording times of observations j and j', and for spatial data the distances would be actual physical distances (*eg* between households). Table 23.2 lists some examples of correlation structures defined from such distances. The structures are **isotropic** when only the distances, not the actual locations for observations j and j', are used. The power (or exponential) structure is the extension of the ar(1) structure to non-equidistant time points; the parameter ρ equals the correlation between 2 observations one unit apart.

Table 23.2 Spatial (or non-equidistant repeated measures) correlation structures

Name	Correlation structure	Interpretation
power, or exponential	$\text{corr}(Y_j, Y_{j'}) = \rho^{d_{jj'}} = \exp(-d_{jj'}/\theta)$	power decay with distance; note the relationship: ρ=exp(-1/θ)
power, or exponential, with nugget effect	$\text{corr}(Y_j, Y_{j'}) = \frac{\sigma^2}{\sigma^2+\sigma_0^2}\rho^{d_{jj'}}$	power decay with distance, close observations not fully correlated
Gaussian	$\text{corr}(Y_j, Y_{j'}) = \exp(-d_{jj'}^2/\theta)$	exponential-quadratic decay with distance
linear	$\text{corr}(Y_j, Y_{j'}) = \begin{cases} 1-\rho d_{jj'} & \text{if } \rho d_{jj'} < 1 \\ 0 & \text{if } \rho d_{jj'} \geq 1 \end{cases}$	linear decay with distance

23.3.2 Linear mixed models with complex correlation structure

Recall that, in the linear mixed model from Chapter 21 (Eq 21.8):

$$Y = X\beta + Zu + \varepsilon \qquad \textit{Eq 23.3}$$

we assumed the components of ε to be independent, and modelled the hierarchical structure using the random effects in the Zu part of the model. In order to enable complex correlation structure, in particular autocorrelation, we will now allow dependence corresponding to a particular correlation structure within some sets of ε-values. In the repeated measures context, each set contains all the repeated measures for a subject, and in the spatial context, each set contains a particular group of observations for which we want to model a spatial correlation (*eg* households within a certain area).

In such mixed models with correlation structure, both the random part Zu and the error correlation structure contributes to total (co)variance (not explained by the fixed effects). If random effects are specified at the same level as the error correlation structure, *eg* subject random effects and within-subject correlation structure, the resulting model may be difficult to estimate and in worst cases even be overparameterised. To illustrate the problem, random effects (intercepts) for subjects cannot be fitted in a model with compound symmetry correlation structure for subjects. This is because both parts of the model will lead to the same correlation structure, so only one of them is needed. Random effects for subjects can however be combined with an ar(1) structure; this produces a structure with autocorrelations that does not decay to zero but instead to the *ICC* one could compute from the between- and within-subject variances. Similarly, a Toeplitz structure cannot be combined with subject random effects, and, if unstructured correlations are specified, it is pointless to include any random effects at the subject level (either a random intercept and random slopes). If the model in Eq 23.3 has no random effects, it is perhaps misleading to call it a mixed model; the name **covariance pattern model** is also used (Hedeker and Gibbons, 2006, Chapter 6).

The statistical analysis of the 'extended' mixed models evolves along the same lines as previously discussed, only with additional variance parameters to be estimated. For large structures (with many time points), parsimonious models for cov(Y) are recommended unless the number of subjects is very large, to avoid overspecification of the model and unexpected impacts of the covariance structure on the fixed-effects parameters. The choice of correlation structure can be formalised by using likelihood-ratio statistics to test nested correlation-

structure models against each other. For example, the compound symmetry and ar(1) models can be tested against both arma(1,1) or Toeplitz models, but the latter two cannot be tested against each other. The test of a compound symmetry model against any of these models allowing for autocorrelation is one of the best ways to assess whether the compound symmetry model is inadequate due to unmodelled autocorrelation. The fit of models with the same number of parameters can be compared by their log-likelihood values (the higher log-likelihood model is generally preferred). Model selection criteria such as the AIC (Chapter 15) are also applicable here. A visual assessment of the residual autocorrelation function may also be helpful (for illustrative examples, see Pinheiro and Bates, 2000, Chapter 5. Example 23.5 examines different correlation structures for the full blood pressure data.

Mixed models with complex correlation structure are currently only available in a few software packages: Stata (Rabe-Hesketh and Skrondal, 2012), SAS (Littell *et al*, 2006), R/S-Plus (Pinheiro and Bates, 2000) and SPSS.

Example 23.5 Linear mixed models with correlation structure for blood pressure data
data = bp

Several correlation structures were examined for the blood pressure measurements at visits 3–6, using additive fixed effects of treatment (-tx-), visit and initial blood pressure (-dbp1c-). While the interaction tx*visit turns out to be non-significant, expansion of fixed effects by interactions with -dbp1c- as well as refinement of the linear term for -dbp1c- could be explored, our present focus is on the random part of the model so we restrict ourselves to the simpler additive form. The model is now a 3-level model with centre random effects and within-patient correlation structure.

Covariance/ correlation structure	Covariance parameters	Estimated ρ 1 visit	2 visits	3 visits	-2 ln likelihood
compound symmetry	1	0.491	0.491	0.491	7470.79
ar(1)	1	0.534	0.285	0.153	7492.69
non-equidistant power	1	0.588	0.345	0.070	7539.73
patient random effects with ar(1) errors	2	0.536	0.447	0.430	7459.96
arma(1,1)	2	0.534	0.463	0.401	7460.86
Toeplitz	3	0.537	0.449	0.425	7459.92

Note The within-patient correlations computed here do not incorporate centre clustering.

The table above illustrates how the different structures adapt to the data. In terms of statistical significance, the Toeplitz model is no better than the hierarchical model with ar(1) errors, which gives the best fit with 2 parameters and is clearly preferable to the structures with only one correlation parameter. The estimated correlations for blood pressure measurements 1, 2, and 3 time steps (visits) apart demonstrates the deficiencies of the one-parameter models. The decay in correlations with distance appears modest, but it is clearly significant. For the non-equidistant structure, visit 6 was assumed to have taken place 6 weeks after visit 5, as described in Hall *et al* (1991), although we have elsewhere (in Chapters 21 and 31) described the post-treatment visits to be equidistant (Brown and Prescott, 2006). The correlations shown in the table are with visit 3, and the very low estimated correlation 3 visits apart is due to the added time gap.

(*continued on next page*)

Example 23.5 (*continued*)

For comparison with the results in previous examples and a subsequent analysis by GEE procedures (with discussion of the parameter estimates), we also present a table of estimates for the fixed effects and random parameters from the hierarchical model with ar(1) errors. The fixed-effects estimates and SEs were very close for the models with comparable fit.

	Coef	SE	z	P	95% CI	
tx = Nifedipine	-1.225	0.974	-1.26	0.209	-3.134	0.684
tx = Atenolol	-2.999	0.964	-3.11	0.002	-4.887	-1.110
visit = 4	-1.000	0.486	-2.06	0.040	-1.952	-0.048
visit = 5	-2.626	0.537	-4.89	<0.001	-3.679	-1.574
visit = 6	-3.093	0.552	-5.60	<0.001	-4.175	-2.011
dbp1c	0.469	0.086	5.47	<0.001	0.301	0.637
constant	94.68	0.89	-	-	92.95	96.42

In addition, the estimated autocorrelation parameter and variance components (also with SEs) were:

$$\rho = 0.193\,(0.061), \quad \text{and} \quad \sigma_c^2 = 4.76\,(2.50), \sigma_p^2 = 30.36\,(4.43), \sigma^2 = 40.98\,(3.00).$$

All the models considered above assume measurements on different patients within a centre to be equally correlated across all visits, but we could imagine correlations to be stronger for measurements taken at the same visit. We can allow for this in our models by adding random effects for centre by visit combinations, or centre by visit by treatment combinations, similar to our modelling in Example 21.5. Such added random effects would break the hierarchical structure, and instead we would have a cross-classification of measurements within both patients and centre by visit combinations which would make the models more difficult to fit. In the blood pressure data, however, very little variance can be explained by centre by visit interactions (results not shown), perhaps due to the short time series.

23.3.3 Trend models

Let us first recap from Section 23.3.1 that trend models are characterised by having random slopes of time. We already argued in favour of these random slopes because they would introduce autocorrelation into the model. It could also be said that models assuming all subjects develop in the same way over time are unrealistic for most longitudinal data (Hedeker and Gibbons, 2006, Chapter 4). With the inclusion of subject-random slopes for time, the model includes terms representing the development over time at the population level (the fixed effects for time) as well as the random effects representing the development over time at the individual level.

We need to be a bit more specific when it comes to how time should be modelled. The simplest option is a linear effect of time but it is often too simplistic to assume a linear change across time; for example, linear trends may eventually level off towards a plateau or a minimum level. The choice of an appropriate form of the time effects follows the same principles as for other continuous predictors (Section 15.4). Ideally, for consistency, the same (linear or non-linear) relation with time would be used for the fixed and random effects. As was noted in Section 21.3, with the need for some parsimony in our use of random slopes, it becomes attractive to consider models with time effects represented by only a few parameters. In some situations, a non-linear monotone transformation (*eg* log or square-root) of the time scale can help to

achieve an approximately linear relation. If the effects of time need multiple parameters, it is helpful for estimation and interpretation of variance parameters if these are as unrelated ('independent') as possible. For example, if a polynomial model is used, it is recommended to parameterise it by orthogonal polynomials (Hedeker and Gibbons, 2006, Chapter 5). If it is not possible to model time effects in a simple fashion, one may choose a simpler form (*eg* linear) for the random slopes while retaining a more complex form for the fixed effects. Modelling time as a categorical random slope predictor (as in Example 21.5) will not introduce autocorrelation into the model and is, therefore, usually less attractive.

For further details about trend models, we refer to the general sections on random slopes (Section 21.3) and inference for mixed models (Section 21.5). In Example 23.6, we give the results of fitting random slopes for time at both the patient and centre levels.

23.4 Mixed models for discrete repeated measures data

From the relative ease with which the linear mixed model could be extended to incorporate autocorrelation, one might expect things to be similar for discrete data, but that is not so. After

Example 23.6 Linear trend models for blood pressure data
data = bp

The random-intercept model (with compound symmetry correlation structure) of Example 23.5 was extended with linear random slopes for visit (centred at visit 3) at both the patient and centre levels. In order to assess how patient-level random slopes would compare with the correlation structures explored in Example 23.5, the first model fit added only patient-level random slopes to the random intercept model. The second model included random slopes at both levels and autocorrelated errors. Only parameter estimates (with SE) for the parameters of the random part of the model are shown.

Level	Parameter	Patient-level random slopes Estimate (SE)	Both level random slopes Estimate (SE)
centre	variance (intercept)	4.788 (2.528)	3.170 (2.605)
	variance (slope)	-	0.384 (0.263)
	covariance	-	0.157 (0.613)
patient	variance (intercept)	39.209 (5.605)	31.885 (5.580)
	variance (slope)	2.005 (0.870)	0.001 (0.012)
	covariance	-2.587 (1.775)	-0.165 (1.255)
measurement	variance	33.041 (2.036)	39.372 (2.835)
	autocorrelation	-	0.167 (0.061)
-2ln likelihood		7464.62	7452.43

By comparison with Example 23.5 we see that for compound symmetry errors, the patient-level random slope improves the correlation structure significantly but not as much as autocorrelation. In the presence of autocorrelation, the patient-level random slope has no real impact but the centre-level random slope is significant. We may interpret this by saying that patients at different centres are on different trajectories over time (in their blood pressure values). The autocorrelation remains significant and is only slightly reduced in magnitude by the added random slopes (from 0.193 to 0.167).

explaining the challenges of adding correlation structure to a GLMM, we give pointers to some of the many different approaches that have been tried, describe in more detail the concept of a **transitional** model, and illustrate by the blood pressure data the impact this approach and the trend model (from the previous section) has on estimates from a random-intercept logistic regression model.

23.4.1 Adding correlation structure to a GLMM

The linear mixed model approach of incorporating correlation structures into the model's error component (ε) runs into the serious problem in GLMMs that the linear predictor (*eg* Eqs 22.1 and 22.5) does not contain an error component. The reason is that a GLM(M) models the mean and variance on different scales: the mean on the scale of the linear predictor given by the link function (in short, the link scale), but the variance on the observation scale. As the error term is on observation scale, it is subject to the restrictions related to the discrete outcome (see *eg* Section 16.1). A second problem following from this is the separation of correlation into parts explained on different scales; recall that in linear mixed-models correlation could be split between the random effects and the error correlation structure (Section 23.3.2), but the situation is more complex when these are on different scales. The third problem is that modelling clustering on the link scale yields parameters with a cluster-specific (SS) interpretation, whereas modelling clustering on observation scale (*eg* in a beta-binomial model, Section 22.4.5, or by the generalised estimating equations, Section 23.5) yields parameters with a population-average (PA) interpretation. If clustering is modelled on both scales, it is not clear which interpretation the parameters will have.

It is therefore more difficult to incorporate the correlation structures discussed for linear mixed models into a GLMM, and this is one of the reasons why no general GLMM-type class of models exists for repeated measures and spatial structures. Instead, models are, to a large extent, developed specifically for the most interesting data types: binary and count data. The literature in this field is large, technical, and largely beyond the scope of this book. We introduce a few of the ideas that tie in with the GLMM framework.

The random-intercept model (Eq 22.1) includes a single random effect for each cluster. As this will not suffice to create a within-cluster correlation structure (Diggle *et al*, 2002, Chapter 11) expanded the model on link scale by including random effects at each time point (for each subject). In a binary model, with probabilities p_{ij} for subject i at time point j, the extension of Eq 22.1 therefore takes the form:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 X_{1ij} + \ldots + \beta_k X_{kij} + u_{ij}, \quad \text{with} \quad u_{ij} \sim N(0, \sigma^2) \qquad \textit{Eq 23.4}$$

The idea is now to assume the set of random effects on each subject, ($u_{i1},\ldots,u_{im}$), to be autocorrelated, *eg* according to the ar(1) structure with correlation ρ (note that ρ is the correlation between the random effects, not between the binary outcomes). In the special case ρ=1, the random effects will be perfectly correlated and thus identical, so that we're back in the random-intercept model with a single random effect (u_i). Unfortunately, the model is difficult to estimate (the MCMC methods of Chapter 24 were suggested as an option). The same random-effects structure has also been applied to autocorrelated count data, *eg* in a times series of counts (Davis *et al*, 2000).

Quasi-likelihood or pseudo-likelihood estimation software may allow specification of a repeated measures or spatial model for the adjusted variate computed in each step of the

iteration (Section 22.5.2). This will lead to correlation structures of repeated measures or spatial type (Gotway and Wolfinger, 2003), although covariance parameters specified in this way may have no direct interpretation in the discrete model. Molenberghs and Verbeke (2005), Chapters 8 and 22) discussed the approach (as implemented in SAS, Proc Glimmix), and it was also included among the methods assessed in recent simulation studies (Masaoud and Stryhn, 2010). Some of the conclusions were: (i) if only a correlation structure is used, the procedure yields estimates with a PA interpretation and is comparable to GEE estimation; (ii) if both random effects and a correlation structure is used, the estimates will be intermediate between PA and SS parameters, and thus be biased for both interpretations.

A GLMM with a correlation modelled at the original scale (Barbosa and Goldstein, 2000) and a multivariate multilevel logistic model (Yang *et al*, 2000) have been developed, but both require specialised software (MLwiN macros), and do not appear to have been used much. A multivariate model for discrete outcomes is also available in MLwiN and can be used for repeated measures data (Rasbash *et al*, 2008); without additional hierarchical structure, the parameters have a PA interpretation.

Repeated measures of counts have been modelled with random effects by a variety of approaches (Nelson and Leroux, 2006), including also the extensions of the Poisson regression reviewed in Chapter 18 (*ie* zero-inflation (Min and Agresti, 2005); overdispersion (Molenberghs *et al*, 2007) and the transitional models to be described next (Li *et al*, 2007)).

23.4.2 Transition models

A generally accepted classification of modelling approaches for clustered data (including repeated measures) is into 3 types: subject-specific, marginal, and transitional (Diggle *et al*, 2002, Chapter 7). Our discussion has until now focused on the first 2, but we will here outline the third approach and explain how it can be used to incorporate autocorrelation into a GLMM. To focus on the basic idea, we consider the simplest case of a binary outcome.

In random-effects models, we accounted for the within-subject clustering by modelling the probability of the event for subject i at time point j conditionally on the (latent) subject random effect u_i, but it might seem more intuitive to model the probability conditionally on the previous event $Y_{i,j\text{-}1}$, and perhaps further events before that. A one-lag transition model (conditioning only on the previous event) could be expressed using the notation of Eq 22.7 as:

$$\text{logit}(p_{ij}) = (X\beta)_{ij} + (Zu)_{ij} + \gamma Y_{i,j-1}, \qquad \textit{Eq 23.5}$$

where only the transitional term $\gamma Y_{i,j\text{-}1}$ is new. Note, it is not a misprint that the outcome Y is present on the right hand side of the equation. The fixed-effect parameter γ equals the log *OR* for a comparison between subjects who at the previous time did and did not experience the event. The model in Eq 23.5 still includes subject random effects because the transitional term cannot be expected to account for all within-subject clustering. Conversely, even if we expect the transitional term to pick up autocorrelation, there may be still be some unmodelled autocorrelation left in the data. A transitional term is sometimes used informally in this way to capture autocorrelation in the data (Thurmond *et al*, 2005).

In Eq 23.5, the probability of an event at time j is different for a preceding non-event ($Y_{i,j\text{-}1}$=0) and a preceding event ($Y_{i,j\text{-}1}$=1), so essentially the model fits an equation for both of these 2 situations. If a disease event occurring after a non-event is interpreted as a new case, the former

situation corresponds to **incidence.** The probability of a disease event following an event would then be interpreted as 1 minus the **cure rate**. In other words, the model in Eq 23.5 is really for the 2 transitions: 0→1 (new case), and 1→0 (cure). With this interpretation, it may seem awkward that the impact of our predictors is assumed to be equal for both transitions; it means that the predictors have numerically exactly opposite effects for incidence and cure. In order to avoid this assumption (which should certainly not be considered the default), we can add interaction terms between $Y_{i,j\text{-}1}$ and the predictors. Similarly, we may want to include a random slope for $Y_{i,j\text{-}1}$. The regression parameters in models such as Eq 23.5 (with or without added interactions) are different than those in the marginal and cluster-specific equations (Eq 22.7), and no general conversion formula exists between them (Diggle *et al*, 2002, Chapter 7). Another issue that distinguishes a transitional model from a usual random-effects model is that a special handling is needed for the first time point (j=1), where no previous outcome is available as a predictor. The values for j=1 may either be omitted, or if they are included, the predictors should contain a dummy variable for this time point and Y_{i0} should be set to zero. In both cases, the model/data will not be quite the same as in the usual random-effects model which will lead to further differences in the parameters. We demonstrate the transitional model in Example 23.7 in the next section.

23.4.3 GLMMs without explicit correlation structure

Although our focus has been on alternatives to the random-intercept model with subject random effects, this simpler model may still be valid, provided one is willing to accept its lack of autocorrelation. To detect violations of compound symmetry may require much more data than in the continuous case because the information content is lower in discrete data— something that certainly is true for binary observations. (**Not**e We may think of the correlation structure as compound symmetry, although strictly speaking the within-subject correlations are only constant when the fixed effects are constant (because the variance is a function of the mean), and in a repeated measures model one would usually have time as a fixed effect.) However, a recent simulation study on binary data (Masaoud and Stryhn, 2010) concluded that even with a repeated measures series as short as m=4, biases may result from ignoring autocorrelation generated by *eg* the model in Eq 23.4. An earlier study cautioned against using the random-intercept model in the presence of autocorrelation (Heagerty and Kurland, 2001) based on a theoretical assessment of the bias in the estimates and a simulation study with m=5 from the same model.

In Example 23.7, we illustrate how the 2 approaches discussed here for modelling correlation structure, random slopes for time (trend models, Section 23.3) and transitional models, affect the estimates of a moderately-sized binary repeated measures dataset. We use again the blood pressure data, and define a binary outcome (-highdbp-) at the threshold of (at least) 90 mmHg, which is in general use to classify diastolic blood pressure values as indicative of hypertension. With only 4 time points (visits) the series is fairly short, so the results may not be representative for longer binary series of repeated measures.

23.5 GENERALISED ESTIMATING EQUATIONS

The previous chapters have presented mixed models as an approach for dealing with the problem of clustering (lack of independence among observations) in a dataset. As noted, these mixed models are very flexible and can handle any number of levels of hierarchical clustering

Example 23.7 Generalised linear mixed models for high blood pressure data
data = bp

Out of the 1,092 blood pressure records, 705 (64.6%) were 90 mmHg or larger and thus indicative of hypertension. Among the 288 patients, 133 (46.2%) had constant values of -highdbp- throughout the study, suggesting, in view of the short series, a moderately strong within-patient clustering. In the next table, we compare estimates from a random-intercept model, a simple transitional model (without any interactions), and a linear-trend model with a patient-level linear random slope for visit centred at visit 3. The effective dataset for the transitional model excluded 8 patients and 287 records where the previous blood pressure value was missing or did not exist (*ie* at visit 3). Note that we don't include the values at visit 3 by conditioning on pre-treatment values because the transition mechanism at treatment onset can hardly be considered as similar as during the follow-up period.

Model	Random intercept	Transitional	Linear trend
Parameter	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
tx = Nifedipine	-0.313 (0.310)	-0.081 (0.265)	-0.479 (0.329)
tx = Atenolol	-0.519 (0.308)	-0.256 (0.264)	-0.619 (0.326)
visit = 4	-0.343 (0.231)	-	-0.218 (0.259)
visit = 5	-1.381 (0.235)	-0.858 (0.229)	-1.210 (0.300)
visit = 6	-1.339 (0.237)	-0.647 (0.253)	-1.101 (0.332)
dbp1c	0.128 (0.029)	0.086 (0.029)	0.148 (0.032)
constant	1.960 (0.309)	0.592 (0.451)	1.887 (0.352)
previous outcome	-	0.962 (0.332)	-
centre variance	0.289 (0.277)	0.202 (0.166)	0.351 (0.311)
patient variance	2.437 (0.548)	0.942 (0.690)	1.515 (1.030)
random slope variance	-	-	0.455 (0.249)
random slope covariance	-	-	0.214 (0.355)

The 2 extensions of the random-intercept model were both highly significant: the coefficient for $Y_{i,j\text{-}1}$ in the transitional model is much larger than its SE, and the linear trend model improved the 2lnL by 9.6 for 2 df. In the transitional model, the odds ratio for having the same outcome as at the previous time point was high ($e^{0.96}$=2.6), and the between-patient variance dropped substantially. The estimates are conspicuously different between the 3 models. This is in part due to the moderately strong within-patient clustering, *eg* in the random-intercept model, the (latent variable) *ICC* equals (0.289+2.437)/(0.289+2.437+3.29)=0.45 (Section 22.2.3). The estimates from random effects models are not directly comparable when the variances are high because of the scaling caused by the random effects (Section 22.6). Estimates become comparable after scaling to PA scale; for example, for -dbp1c- we obtain:

$$0.128/\sqrt{1+0.346*(0.289+2.437)}=0.092, \quad \text{and} \quad 0.086/\sqrt{1+0.346*(0.202+0.942)}=0.073,$$

so these estimates are not quite as far off as it appears. The estimated intercept in the transitional model has a different interpretation and role in the model: it corresponds to tests whose predecessor was zero, whereas in the other models, it corresponds to any measurement. Due to non-constant variances created by the random slopes (the estimates here imply a marked increased in variance at later visits), it is more difficult to scale the estimates of the trend model. The conclusion from the example is that both extensions of the random-intercept model affect the model quite strongly, and it is not clear which (if any) of them is preferable. We continue the analysis of this model in Example 23.10.

as well as more complex data structures. However, some unresolved issues remain. As discussed in Section 23.4, the mixed model approach is not as successful with repeated and spatial structures for discrete data as it is for continuous data. Also, its assumption of normally distributed random effects is perhaps a limitation; in practice, you will encounter data that clearly do not conform to that assumption. From a more philosophical point of view, one might argue that, in our analyses, we should only make the absolutely necessary distributional assumptions and for 'nuisance effects', rely on robust procedures that are less affected by the peculiarities of the data. This would follow the trend in modern statistics toward non- and semi-parametric procedures, as seen, for example, in survival analysis. Finally, complex mixed models are sometimes difficult to fit due to the size of the data or to numerical difficulties.

Generalised estimating equations were introduced in 2 papers by Liang and Zeger (1986) and Zeger and Liang (1986) as a set of estimating equations to obtain parameter estimates for discrete and continuous repeated measures data. The idea has proven not only durable but also extendable to other data structures (*eg* hierarchically clustered and spatial data), statistical inference accompanying the estimates, as well as many variants of estimating equations (Hanley *et al*, 2003). A fairly recent (statistical) monograph (Hardin and Hilbe, 2003) is devoted entirely to GEE methods, which today are one of the most popular approaches in the health and biological sciences. We will confine ourselves here to describing the original (and probably still most popular) GEE method to obtain population-averaged estimates for clustered data. To illustrate the methods, we will use the repeated measures blood pressure data and the hierarchically clustered diarrhea data from Chapter 22.

23.5.1 Estimating equations

Let's initially explain the meaning of an 'estimating equation'. When using maximum likelihood (ML) estimation, the parameters are chosen to maximise the log-likelihood function. In practice, maximising a function involves computing the (partial) derivatives of the function with respect to its parameters and equating these to zero. These would be the estimating equations for ML estimation (and the derivatives of the log-likelihood function is called the score function). Except for very simple cases, the equations do not have an explicit solution and must be solved iteratively. The approach we are going to take here involves GLMs and a partially specified model, so that no likelihood function is available. Specifically, the GEE method requires only assumptions about the marginal mean and variance (and information about the subjects, or more generally clusters, of the data). Nevertheless, estimation is based on iterative solution of similar generalised estimating equations. These equations involve the mean of the outcome across clusters, therefore GEE yields estimates with a **PA interpretation**. Recall however from Section 22.4.1, that a distinction between SS and PA estimates is unnecessary for models with an identify link, such as a linear (mixed) regression models.

23.5.2 Statistical inference using GEE

The Liang and Zeger version of GEE is based on correlations in a working correlation matrix. Despite the fact that no assumptions about the form of the correlation of the data within the clusters are made, the estimating equations involve a **working correlation matrix** containing the estimated correlations among observations within a cluster, in each cycle of the iterations. This matrix can be given different forms (independent, compound symmetry, autoregressive, unstructured *etc* as in Section 23.3.1) to tailor the estimating algorithm toward one's perception

of the data structure. Because the matrix is not part of the model, its form is not as crucial as in a fully parametric model. Theoretically, the GEE method gives asymptotically unbiased estimates even if the working correlation matrix is misspecified; that might, however, lead to loss in efficiency (Fitzmaurice, 1995). Estimation of variance (*ie* standard errors and correlations among estimates) can be either model-based or robust (or empirical) as described in Section 20.5.4. The latter method is also asymptotically unbiased, and is generally recommended because the GEE method loses its robustness to misspecification if model-based variance estimation is used. It is worth noting the general relationship that GEE with independent working correlation structure and robust variance is exactly the same as ordinary clustered robust variance estimation (Section 20.5.4).

As to the choice of working correlation structure, you should first and foremost be guided by your understanding of the data. For hierarchically clustered data (*eg* individuals in families), anything but a compound symmetry (or exchangeable) correlation structure would seem unreasonable. Particular caution should be exercised with negatively correlated binary data. In this case, an ordinary logistic model with robust standard errors has been recommended (Hanley *et al*, 2000). For repeated measures data, one would usually choose a structure that allows for autocorrelation. It might also be tempting to try an unstructured correlation to see what patterns the data show when not constrained by a particular structure. However, large correlation structures imply estimation of a large number of 'working parameters' and numerical problems might be encountered especially in unbalanced datasets. Recently a criterion (QIC), similar to Akaike's information criteria has been developed to guide the choice of correlation matrix (Pan, 2001) and implemented in standard software (Cui and Qian, 2007). We first illustrate the GEE method in Example 23.8 by applying it to the multilevel diarrhea data with a binary outcome.

A word of caution about the use of the GEE approach is appropriate when it comes to missing data (Section 15.5). It has long been recognised that GEE is not robust to missing data under the missing at random (MAR) assumption, but that addition of a weighting scheme to the procedure could resolve the problem (Robins *et al*, 1995; Molenberghs *et al*, 2007, Chapter 27). The actual scheme depends on the structure of the missing values (*eg* whether these are drop-outs or intermediate missing values). An implementation of a weighting scheme for drop-outs has been published (Jansen *et al*, 2006), but such adjustments to GEE do not seem to be generally available in standard statistical software.

23.5.3 GEE for multilevel data structures

One apparent drawback of the GEE method is its limitation to a single level of clustering. Except for the alternating logistic regression (ALR) version of GEE discussed below, the problem of extending GEE algorithms to account for more complex data structures has received relatively little attention in the literature (Chao, 2006; Teerenstra *et al*, 2010). The question of how to best set up a classical GEE analysis for binary repeated measures with an added hierarchical level (*eg* patients clustered in centres) was discussed on the basis of multiple simulation studies (Masaoud, 2009). The recommendations were that, with moderate-to-large numbers of highest level clusters, it is sufficient to cluster at the highest level to achieve approximately unbiased estimates and standard errors at all levels, and that other schemes such as ignoring the highest level clusters or modelling them by fixed effects were less successful. This finding agrees with the recommendation by Hardin and Hilbe (2003), Chapter 3 that for complete datasets with a number of clusters above 30, there is little gain in using more

Example 23.8 Generalised estimating equations for family-level diarrhea data
data = brazil_smpl

For the simple family-level model of Example 22.12, a GEE analysis with a compound symmetry structure within municipalities and robust standard errors gave a regression coefficient for cistern of: -0.612 (0.221). For comparison with the corresponding random-effects estimate (-0.668), we compute its PA counterpart using Eq 22.2: $\beta^{PA} \approx -0.668/\sqrt{1+0.346*0.469} = -0.620$. Thus, the 2 estimates agree very closely; however, the SE is appreciably larger for the GEE estimate even before the SE of the GLMM estimate (0.162) is scaled towards zero. The most likely explanation for the disagreement is that the robust variance estimation included in GEE ideally requires at least 30 clusters to perform satisfactorily (and we only had 21 municipalities); indeed, the robust SE is also elevated for the random effects model (0.244). The discrepancy between the SEs is smaller when adjusting for clustering at communities instead (GEE estimate: -0.629 (0.183), which can be compared with the values of Example 22.1). When accounting for both levels of clustering by the ALR method, the SE (0.217) is intermediate between these two values but still markedly higher than from the GLMM.

The municipality-level working correlation matrix had a correlation of 0.073, which is substantially lower than the approximate *ICC* computed by the latent variable method as 0.469/(0.469+3.29)=0.125. This disagreement is not related to cluster size and thus perhaps reflects the different ways of estimating correlation by the two approaches.

complicated correlation structures than independence (which de facto is ordinary logistic regression with robust standard errors), and it also agrees with the common approach to survey data to adjust for clustering by primary sampling units and pay less attention to subsequent levels (Sections 2.10 and 20.5.5). As a higher-level working correlation structure cannot easily be set up to account for autocorrelation within subjects, an exchangeable structure must be used and a possible loss of power by misspecification of the structure must be accepted. In Example 23.9, we compare different 3-level GEE approaches to analysis of a continuous outcome by reanalysing the blood pressure data from Examples 23.5–6.

For binary outcomes, an alternative to the standard GEE algorithms was developed by Carey *et al* (1993) and termed **alternating logistic regression** (ALR) because the estimation algorithm in each step of the iterations employs 2 (very different) logistic regression models to update the parameters. As this approach was favoured for binary data in a comprehensive review of GEE methods (Hardin and Hilbe, 2003), and it also has the ability to deal with 2 levels of clustering, we briefly describe the idea and demonstrate in Example 23.10 its use (together with other GEE implementations) to the high blood pressure model of Example 23.7. The standard GEE procedure describes within-subject clustering in terms of a working correlation matrix; however, correlation is not the most obvious measure of association for binary outcomes. The ALR method instead describes the clustering in terms of odds-ratios for 2 subjects within the same cluster, and offers estimates (with SEs) of such quantities. As the estimating equation for the fixed effects is the same as for standard GEE, the robustness properties of GEE are retained. One drawback of the approach is that it is only implemented in a few statistical packages (SAS and R/S-plus) and only with exchangeable correlation structures. That is, in the repeated measures context, the odds-ratio parameter gives the ratio between odds of disease when it is known that another observation on the same subject is disease-positive versus when it is disease-negative. A numerical illustration is given also in Example 23.10.

Example 23.9 Generalised estimating equations for blood pressure data
data = bp

We analysed these data using linear mixed models for repeated measures in Examples 23.5–6. Because of the identity link function, the SS and PA parameters coincide. The difference of the GEE approach lies therefore entirely in the estimation method. The table shows parameter estimates from GEE analyses clustered at the patient level with compound symmetry, autoregressive (ar(1)), and unstructured working correlation matrices. The table also gives values of the working correlations 1, 2, and 3 time steps apart; the values for the unstructured correlation were obtained by averaging the corresponding values in the matrix. Some software implementations of GEE (*eg* in SAS) will fit time-dependent correlation structures without excluding incomplete sets of repeated measures. Although this is generally preferable, for simplicity and in order to enable meaningful comparisons, the results here are without the 8 patients that only appeared at visit 3 and with gaps ignored for 1 patient.

	Patient-level working correlation matrix structure			Centre work. corr.
Model	**Comp. symm.**	**Autoregressive**	**Unstructured**	**Comp. symm.**
Parameter/Statistic	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
tx = Nifedipine	-1.258 (0.999)	-1.158 (1.006)	-1.084 (0.993)	-1.233 (0.945)
tx = Atenolol	-2.958 (1.075)	-2.918 (1.070)	-2.803 (1.062)	-2.989 (0.974)
visit = 4	-1.084 (0.504)	-1.068 (0.505)	-1.076 (0.505)	-1.078 (0.446)
visit = 5	-2.740 (0.556)	-2.702 (0.559)	-2.717 (0.557)	-2.784 (0.812)
visit = 6	-3.221 (0.553)	-3.206 (0.554)	-3.210 (0.552)	-3.282 (0.900)
dbp1c	0.507 (0.108)	0.489 (0.109)	0.499 (0.106)	0.480 (0.120)
constant	94.19 (0.80)	94.15 (0.81)	94.08 (0.80)	95.20 (0.83)
ρ (1 visit)	0.508	0.557	0.559	0.131*
ρ (2 visits)	0.508	0.310	0.476	0.131*
ρ (3 visits)	0.508	0.173	0.417	0.131*
QIC	82005.78	82012.55	82007.53	83256.61

*correlation among all values within a centre

The estimates are in reasonable agreement with those of the linear mixed model in Example 23.5 (the comparison is made difficult by the slight differences in datasets). The GEE estimates also agree fairly well between methods for treatments and the covariate (dbp1c), and are very close for time (visits). The standard errors are almost identical for the 3 within-patient correlation structures, but substantially off when adjusting for clustering at centres. These results demonstrate that choice of working correlation structure is not always of minor importance for the fixed effects, even in a moderately sized dataset. The QIC points, perhaps surprisingly, to the compound symmetry structure as the preferable one, despite the obvious autocorrelation in the unstructured working correlation matrix, and the analysis clustered at centres is firmly rejected. Thus, model choice by the QIC statistic leads to different conclusions than in the mixed model and the recommendations from simulation studies.

In summary, there is a fair agreement between the GEE and linear mixed models analysis, but some questions remain for GEE with respect to choice of working correlation structure and the handling of incomplete series. Both analyses showed a clearly significant difference between Carvedilol and Atenolol (the former reducing blood pressures the least) and a non-significant tendency in the same direction for Nifepidine. As no clear tx by visit interaction was present, the analysis restricted to visits 3–6 with the initial blood pressure as a covariate lead to more clear-cut interpretations than could be obtained from a joint analysis of data from all 5 visits (Example 23.4).

Example 23.10 GEE and ALR estimation for high blood pressure data
data = bp

Using the same dataset as for Example 23.7, we compare different versions of GEE to account for the repeated measures and additional clustering in centres. For comparison, an ordinary logistic regression is included as well. Incomplete series were handled in the same way as in Example 23.9.

Model	Ord. logistic regression	GEE: ar(1) at patients	GEE: cs at centres	Altern. logistic regression
Parameter	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
tx = Nifedipine	-0.237 (0.165)	-0.215 (0.221)	-0.250 (0.206)	-0.231 (0.198)
tx = Atenolol	-0.390 (0.163)	-0.394 (0.227)	-0.407 (0.189)	-0.374 (0.182)
visit = 4	-0.266 (0.196)	-0.268 (0.167)	-0.268 (0.191)	-0.258 (0.183)
visit = 5	-0.990 (0.191)	-0.983 (0.162)	-0.996 (0.346)	-0.975 (0.333)
visit = 6	-0.959 (0.192)	-0.956 (0.175)	-0.961 (0.392)	-0.947 (0.382)
dbp1c	0.097 (0.015)	0.095 (0.019)	0.091 (0.019)	0.091 (0.019)
constant	1.341 (0.174)	1.336 (0.190)	1.414 (0.256)	1.350 (0.240)

We see immediately that the estimates show less variation across different analyses than in Example 23.7. Two explanations can be offered: contrary to Example 23.7, the analyses correspond to the same models (fixed effects), and all the estimates are on the same (PA) scale. One major discrepancy remains regarding the SEs of some estimates where the robust standard errors that account for centre clustering produce much larger SEs. As already discussed in Example 23.7, this may be due to a relatively low number of upper level clusters (29 centres). The 2 ALR log-odds ratio parameters were estimated at:

$$\text{within-patient: } \phi = 1.342\ (0.195), \quad \text{and} \quad \text{between-patient (within centre): } \phi = 0.045\ (.051)$$

The interpretation of these values is that the odds of a high blood pressure at one test is $e^{1.342}=3.8$ times higher when it is known that another test of the same patient was positive than when that other test was negative. By comparison, the odds is only $e^{0.045}=1.05$ times higher when the same information is given about another patient at the same centre, indicating that only minimal centre clustering is present in the data (we reached the same conclusion in Example 23.7). The random-intercept model estimates from Example 23.7 agree well with the present estimates after rescaling to PA scale, *eg* the previously computed PA value of 0.092 for the initial blood pressure (-dbp1c-).

For repeated measures with a short series and few missing values, the GEE procedures are relatively robust to the specification of correlation structure and thus appear as an attractive alternative to random effects models, which involve modelling choices with greater impact on fixed effects inference, as illustrated in Example 23.7.

23.5.4 Summary remarks on GEE and discrete mixed models

We expand here a bit on the summary Table 20.4 to specifically address the choice between GEE and discrete mixed models. The advantage of the GEE method (and many of its generalisations) is that it has robust theoretical properties with few model assumptions. It is also computationally feasible for large datasets and can be fit with a wide range of working correlation structures. It is one of the few general methods for use with discrete repeated measures and spatial data; however, it does not provide much information about the random structure of the data, and it cannot be used to model random structure in terms of random slopes. Its lack of likelihood-based inference and standard errors for correlation parameters are

perhaps less of an issue, but GEE estimation may require additional analysis for data with a large proportion of missing values that cannot be assumed missing completely at random.

A general GLM(M) class of random-effects models that allow inclusion of autocorrelation and other complex correlation structures does not exist (disregarding the quasi-likelihood approach discussed in Section 23.4.1), but a range of specific methods are available for binary and count data. The choice between methods may require a considerable effort to understand their theoretical basis, and can also be difficult in practice, *eg* in binary data with strong within-subject clustering, as indicated in our examples. It is recommended to try multiple approaches in order to assess the robustness of the results to the particular choice of method. Modelling of time by random slopes (trend models) should probably be included among the methods used, unless the time series is very short. The ability to include additional hierarchical structure remains one of the main advantages of mixed models.

References

Barbosa MF, Goldstein H. Discrete response multilevel models. Quality and Quantity. 2000;34:323-30.

Brown H, Prescott RI. Applied Mixed Models in Medicine 2nd Ed. Chichester: Wiley; 2006.

Carey V, Zeger SL, Diggle PJ. Modelling multivariate binary data with alternating logistic regressions. Biometrika. 1993;80:517-26.

Chao EC. Structured correlation in models for clustered data. Stat Med. 2006 Jul 30;25(14):2450-68.

Cui J, Qian G. Selection of working correlation structure and best model in GEE analyses of longitudinal data. Communications in Statistics- Simulation and Computation. 2007;36:987-96.

Davis CR. Statistical Methods for the Analysis of Repeated Measurements. New York: Springer; 2002.

Davis RD, Dunsmuir WT, Wang Y. On autocorrelation in a Poisson regression model. Biometrika. 2000;87:491-505.

Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data, 2nd Ed. Oxford: Oxford University Press; 2002.

Everitt BS. The analysis of repeated measures: A practical review with examples. J R Stat Soc D (The Statistician). 1995;44:113-35.

Fitzmaurice GB. A caveat concerning independence estimating equations with multivariate binary data. Biometrics. 1995;51:309-17.

Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. New York: Wiley; 2004.

Gotway CA, Wolfinger RD. Spatial prediction of counts and rates. Stat Med. 2003;22:1415-32.

Hall S, Prescott RI, Hallman RJ, Dixon S, Harvey RE, Ball SG. A comparative study of Carvedilol, slow-release Nifedipine, and Atenolol in the management of essential

hypertension. J Cardiovasc Pharmacol. 1991;18 Suppl 4:S35-8.

Hanley JA, Negassa A, Edwardes MD. GEE: Analysis of negatively correlated binary responses: a caution. Stat Med. 2000;19:715-22.

Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations; an orientation. Am J Epidemiology. 2003;157:364-75.

Hardin JW, Hilbe JM. Generalized estimating equations. Boca Raton: Chapman & Hall/ CRC; 2003.

Heagerty PJ, Kurland FK. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika. 2001;88:973-85.

Hedeker D, Gibbons RD. Longitudinal Data Analysis: Wiley; 2006.

Jansen I, Beunckens C, Molenberghs G, Verbeke G, Mallinckrodft C. Analyzing incomplete discrete longitudinal clinical trial data. Statistical Science. 2006;21:52-69.

Li J, Yang X, Wu Y, Shoptaw S. A random-effects Markov transition model for Poisson-distributed repeated measures with non-ignorable missing values. Stat Med. 2007 May 30;26(12):2519-32.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13-22.

Littell RC, Milliken GA, Stroup WS, Wolfinger RD, Schabenberger O. SAS for Mixed Models, 2nd ED: SAS Publishing; 2006.

Masaoud E. Statistical models for binary repeated measures and hierarchical data in veterinary science. Charlottetown: University of PEI; 2009.

Masaoud E, Stryhn H. A simulation study to assess statistical methods for binary repeated measures data. Prev Vet Med. 2010 Feb 1;93(2-3):81-97.

Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. Statistical Modelling. 2005;5:1-19.

Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New York: Springer; 2005.

Molenberghs G, Verbeke G, Demetrio CGB. An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Analysis. 2007;13:513-31.

Nelson KP, Leroux BG. Statistical models for autocorrelated count data. Stat Med. 2006;25:1413-30.

Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001;57:120-5.

Pinheiro JC, Bates DM. Mixed-effects Models in S and S-Plus: Springer; 2000.

Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling using Stata, 2nd ED: Stata Press; 2012.

Rasbash J, Steele F, Browne W, Goldstein H. A User's Guide to MLwiN. Centre for Multilevel

Modelling, Bristol: University of Bristol; 2008.

Robins JM, Rotnizsky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Amer Statist Assoc. 1995;90:106-21.

Senn S, Stevens L, Chaturvedi N. Tutorial in biostatistics: Repeated measures in clinical trials: simple strategies for analysis using summary measures. Stat Med. 2000;19:861-77.

Teerenstra S, Lu B, Preisser JS, van Achterberg T, Borm GF. Sample size considerations for GEE analyses of three-level cluster randomized trials. Biometrics. 2010 Dec;66(4):1230-7.

Thurmond MC, Branscum AJ, Johnson WO, Bedrick EJ, Hanson TE. Predicting the probability of abortion in dairy cows: a hierarchical Bayesian logistic-survival model using sequential pregnancy data. Prev Vet Med. 2005 May 10;68(2-4):223-39.

Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data: Springer; 2001.

Yang M, Heath A, Goldstein H. Multilevel models for binary outcomes: attitudes and vote over the electoral cycle. J R Stat Soc A. 2000;163:49-62.

Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42:121-30.