Statistical Methods in Medical Research

P. Armitage
MA, PhD
Emeritus Professor of Applied Statistics
University of Oxford

G. Berry
MA, PhD
Professor in Epidemiology and Biostatistics
University of Sydney

J.N.S. Matthews
MA, PhD
Professor of Medical Statistics
University of Newcastle upon Tyne

FOURTH EDITION

Blackwell
Science

studied and where attention very often focuses on the main effects and low-order interactions.

A further way in which a full factorial design can be reduced, in a block experiment, is to arrange that each block contains only a selection of the possible factor combinations. The design is chosen to ensure that some effects, typically main effects and low-order interactions, can be estimated from contrasts *within blocks*, whereas others (of less interest) are estimated from contrasts *between blocks*. The latter are said to be *confounded with blocks*, and are, of course, estimated with lower precision than the unconfounded effects.

9.6 **Split-unit designs**

In a factorial design in which confounding with blocks takes place, as outlined at the end of §9.5, two types of random variation are important: the variation between experimental units within a block, and that between blocks. In some simple factorial designs it is convenient to recognize two such forms of experimental unit, one of which is a subdivision of the other, and to arrange that the levels of some factors are spread across the larger units, while levels of other factors are spread across the smaller units within the larger ones.

This principle was first exploited in agricultural experiments, where the designs are called *split-plot designs*. In some field experiments it is convenient to divide the field into 'main plots' and to compare the levels of one factor—say, the addition of different soil organisms—by allocating them at random to the main plots. At the same time each main plot is divided into a number of 'subplots', and the levels of some other factor—say, different fertilizers—are allocated at random to subplots within a main plot, exactly as in a randomized block experiment. The comparison of fertilizers would be subject to the random variation between subplots, which would be likely to be less than the variation between main plots, which affects organism comparisons. The organisms are thus compared less precisely than the fertilizers. This inequality of precision is likely to be accepted because of the convenience of being able to spread organisms over relatively large areas of ground.

Similar situations arise in medical and other types of biological experimentation. In general the experimental units are not referred to as 'plots', and the design is therefore more appropriately called a *split-unit design*. Another term is *nested design*. If the subunits are serial measurements on the main units then a split-unit analysis is sometimes called a *repeated measures analysis of variance*: for a discussion of some special considerations that apply in this case, see §12.6.

Some examples of the distinction between main units and subunits are as follows:

Main unit	Subunit
Individual human subject or animal	Different occasions with the same subject or animal
Litter	Animals within a litter
Day	Periods during a day

In the first of these instances a split-unit design might be employed to compare the long-term effects of drugs A_1, A_2 and A_3, and simultaneously the short-term effects of drugs B_1, B_2 and B_3. Suppose there are 12 subjects, each of whom must receive one of A_1, A_2 and A_3; and each subject is observed for three periods during which B_1, B_2 and B_3 are to be given in a random order. The design, determine by randomly allocating the As to the different subjects and the Bs to the period within subjects, might be as follows.

Patient	'A' drug throughout	'B' drug during period		
		1	2	3
1	A_3	B_1	B_3	B_2
2	A_1	B_1	B_2	B_3
3	A_1	B_3	B_1	B_2
4	A_2	B_3	B_2	B_1
5	A_3	B_2	B_3	B_1
6	A_2	B_2	B_1	B_3
7	A_1	B_1	B_2	B_3
8	A_3	B_3	B_1	B_2
9	A_3	B_1	B_3	B_2
10	A_2	B_2	B_1	B_3
11	A_1	B_2	B_1	B_3
12	A_2	B_2	B_1	B_3

The analysis of such designs is illustrated in Example 9.5, using data from a survey rather than an experiment.

Example 9.5

The data in Table 9.10 are taken from a survey on the prevalence of upper respiratory tract infection. The variable to be analysed is the number of swabs positive for *Pneumococcus* during a certain period. Observations were made on 18 families, each consisting of a father, a mother and three children, the youngest of whom was always a preschool child. The children are numbered 1, 2 and 3 in descending order of age. Six families were a

random selection of such families living in 'overcrowded' conditions, six were in 'crowded' conditions and six were in 'uncrowded' conditions.

The first point to notice is that two types of random variation are relevant: that between families (the main units in this example) and that between people within families (the subunits). Comparisons between degrees of crowding must be made *between families*, comparisons of family status are made *within families*. With designs of any complexity it is a good idea to start the analysis by subdividing the degrees of freedom. The result is shown in the DF column of Table 9.11. The total DF are 89, since there are 90 observations. These are split (as in a one-way analysis of variance) into 17 (= 18 – 1) between families and 72 (= 18 × 4) within families. The between-families DF are split (again as in a one-way analysis) into 2 (= 3 – 1) for degrees of crowding and 15 (= 3 × 5) for residual variation within crowding categories. The within-families DF are split into 4 (= 5 – 1) for

Table 9.10 Numbers of swabs positive for *Pneumococcus* during fixed periods.

		Family status					
				Child			
Crowding category	Family serial number	Father	Mother	1	2	3	Total
Overcrowded	1	5	7	6	25	19	62
	2	11	8	11	33	35	98
	3	3	12	19	6	21	61
	4	3	19	12	17	17	68
	5	10	9	15	11	17	62
	6	9	0	6	9	5	29
		41	55	69	101	114	380
Crowded	7	11	7	7	15	13	53
	8	10	5	8	13	17	53
	9	5	4	3	18	10	40
	10	1	9	4	16	8	38
	11	5	5	10	16	20	56
	12	7	3	13	17	18	58
		39	33	45	95	86	298
Uncrowded	13	6	3	5	7	3	24
	14	9	6	6	14	10	45
	15	2	2	6	15	8	33
	16	0	2	10	16	21	49
	17	3	2	0	3	14	22
	18	6	2	4	7	20	39
		26	17	31	62	76	212
Total		106	105	145	258	276	890

categories of family status, 8 (= 4 × 2) for the interaction between the two main effects, and 60 for within-families residual variation. The latter number can be obtained by subtraction (60 = 72 − 4 − 8) or by regarding this source of variation as an interaction between the between-families residual variation and the status factor (60 = 15 × 4). It may be wondered why the interaction between status and crowding is designated as within families when one main effect is between and the other is within families. The reason is that this interaction measures the extent to which the status differences, which are within families, vary from one degree of crowding to another; it is therefore based entirely on within-families contrasts.

Table 9.11 Analysis of variance for data in Table 9.10.

				VR against:	
	SSq	DF	MSq	*a*	*b*
Between families	1146·09	17			
Crowding	470·49	2	235·24		5·22*
Residual	675·60	15	45·04[b]	1·78	1·00
Within families	3122·80	72			
Status	1533·67	4	383·42	15·17**	
Status × crowding	72·40	8	9·05	0·36	
Residual	1516·73	60	25·28[a]	1·00	
Total	4268·89	89			

*$P = 0{\cdot}019$.

**$P < 0{\cdot}001$.

The calculation of sums of squares follows familiar lines. Thus,

Correction Term CT = $(890)^2/90$ = 8801·11

Total SSq = $5^2 + 7^2 + \ldots + 20^2 - \text{CT}$ = 4268·89

Between-Families SSq = $(62^2 + \ldots + 39^2)/5 - \text{CT}$ = 1146·09

Within-Families SSq = Total SSq − Between Families SSq = 3122·80

Subdividing the Between-Families SSq,

Crowding SSq = $(380^2 + 298^2 + 212^2)/30 - \text{CT}$ = 470·49

Residual = Between-Families SSq − Crowding SSq = 675·60

Subdividing the Within-Families SSq,

Status SSq = $(106^2 + \ldots + 276^2)/18 - \text{CT}$ = 1533·67

S × C SSq = $(41^2 + \ldots + 76^2)/6 - \text{CT}$ − Status SSq − Crowding SSq = 72·40

Residual = Within-Families SSq − Status SSq − S × C SSq = 1516·73

The variance ratios against the Within-Families Residual MSq show that differences due to status are highly significant: we return to these below. The interaction is not

significant; there is therefore no evidence that the relative effects of family status vary from one crowding group to another. The variance ratio of 1·78 between the two residuals is just on the borderline of significance at the 5% level. But we should expect a priori that the between-families residual variance would be greater than that within families, and we must certainly test the main effect for crowding against the between-families residual. The variance ratio, 5·22, is significant.

The means for the different members of the family are:

		Child		
F	M	1	2	3
5·6	5·8	8·1	14·3	15·3

The standard error of the difference between two means is $\sqrt{[2(25{\cdot}28)/18]} = 1{\cdot}68$. There are clearly no significant differences between the father, mother and eldest child, but the two youngest children have significantly higher means than the other members of the family.

The means for the different levels of crowding are:

Overcrowded	Crowded	Uncrowded
12·7	9·9	7·1

The standard error of the difference between two means is now $\sqrt{[2(45{\cdot}04)/30]} = 1{\cdot}73$. There is some evidence of a difference between overcrowded and uncrowded families. However, there seems to be a trend and it might be useful to divide the two degrees of freedom for crowding into one for a linear trend and one for the remaining variation (see §8.4).

Split-unit designs more elaborate than the design described above may be useful. For example, the structure imposed on the main units (which in Example 9.5 was a simple one-way classification) could be a randomized block design or something more complex. The subunit section of the analysis would then be correspondingly enlarged by isolation of the appropriate interactions. Similarly, the subunit structure could be elaborated. Another direction of generalization is in the provision of more than two levels in the hierarchy of nested units. In a study similar to that of Example 9.5, for instance, there might have been several periods of observation for each individual, during which different treatments were administered. There would then be a third section in the analysis, within individuals, with its corresponding residual mean square.

The split-unit design, with its two levels of residual variation, can be regarded as the prototype for multilevel models, a flexible and widely used class of models which will be discussed in §12.5.

The following example illustrates a case in which there are two levels of nested units, but in which the design is very simple. There are no structural factors, the purpose of the analysis being merely to estimate the components of random variation.

To be fair to methods which apply transformations directly to data from studies of Michaelis–Menten kinetics, there are several different linearizing transformations, most of which are superior to the Lineweaver–Burk method. One of the best of these is obtained by noting that (12.27) can be written as:

$$\frac{s}{v} = \frac{K}{V_{\max}} + \frac{1}{V_{\max}} s,$$

so a linear regression of the ratios s/v against s has slope $1/V_{\max}$ and intercept $K/V_{\max}$. An extended discussion of several methods for analysing data from Michaelis–Menten studies can be found in Cornish-Bowden (1995a, b).

In some cases linearization techniques effect a valuable simplification without causing any new problems. An example of this is periodic regression, where the response exhibits a cyclic or seasonal nature. A model which comes to mind would be

$$\mathrm{E}(y|x) = \alpha_0 + \alpha_1 \sin(\beta x + \gamma). \tag{12.32}$$

While a cyclic trend cannot always be captured by a simple sinusoidal curve, it can be a useful alternative to a null hypothesis that no cyclical trend exists; a fuller discussion is given by Bliss (1958). An example of the use of this model is given by Edwards (1961), who tested a series of counts made at equally spaced intervals for periodicity. Suppose there are k counts, N_i. For example, neurological episodes may be classified by the hour of the day ($k = 24$), or congenital abnormalities by the month of the year ($k = 12$).

In (12.32) α_0 is the mean level about which the N_i fluctuate, α_1 is the amplitude of the variation, β determines the period of the variation and γ is the phase. If the equation (12.32) is to have a complete cycle of k time intervals then $\beta = 360/k$ (degrees). Even though this deals with one non-linear parameter, the resulting equation is still non-linear because γ does not appear linearly. However, expanding the sine function gives the alternative formula

$$\mathrm{E}(y|x) = \alpha_0 + \zeta_1 \sin(\beta x) + \zeta_2 \cos(\beta x),$$

where $\zeta_1 = \alpha_1 \cos\gamma$ and $\zeta_2 = \alpha_1 \sin\gamma$. This equation is linear in these parameters and the regression can be fitted by recalling that β is known and noting that x successively takes the values $1, 2, \ldots, k$.

12.5 Multilevel models

In the models discussed so far in this chapter the primary concern has been to allow the mean of the distribution of an outcome, y, to be described in terms of covariates x. A secondary matter, which has been alluded to in §12.3 but not discussed in detail, is the modelling of the variance of y in terms of covariates. In all cases it is assumed that separate observations are quite independent of one

another. The distributions of different ys may be similar because the corresponding xs are similar, but knowledge of one of the ys will provide no further information about the value of the others.

However, many circumstances encountered in medical research give rise to data in which this level of independence does not obtain and a valid analysis requires richer models than those considered hitherto. For example, it is quite reasonable to assume that observations on the blood pressure of different patients are independent, but it may well be quite unreasonable to make the same assumption about measurements made on the same patient. This could be because the patient may have a familial tendency to hypertension and so usually has a high blood pressure. Consequently the value on one occasion will give information about subsequent observations. To analyse a series of such observations as if they were independent would lead to bias: e.g. the serial dependence would give rise to an inappropriately small estimate of variance. Another example would be the level of glycaemic control amongst patients with Type II diabetes attending a particular general practice. All patients with this disease in the practice will receive advice on managing their condition from a practice nurse or one of the doctors—that is, the patients share a common source for their advice. If the advisers are particularly good, or particularly bad, then all patients in the practice will tend to benefit or suffer as a consequence.

In these examples the dependence between observations has arisen because the measurements at the most basic level, the individual measurements of blood pressure or the glycaemic control of an individual patient, occur within groups, the individual patient or the general practice, and such data are often referred to as *hierarchical*. It is common for dependence to arise by this means and the class of *multilevel models* provides a family of models that can address the statistical problems posed by this kind of data. In this section only multilevel models for a continuous outcome will be considered, but they can be very usefully employed to analyse other types of outcome. A full discussion of this family of models can be found in Goldstein (1995).

Hierarchical data formed by the serial measurement of a quantity on an individual, also known as *longitudinal data*, occur throughout medicine and can be addressed by a wide variety of methods in addition to those provided by multilevel models. Consequently, discussion of this kind of data is deferred until §§12·6 and 12·7. However, it should be borne in mind that the methods described in this section can often be used fruitfully in the study of longitudinal data.

Random effects and building multilevel models

Rather than attempting to model variances and correlations explicitly, multilevel models make extensive use of random effects in order to generate a wide variety of dependence structures. A simple illustration of this can be found in Example

9.5. The number of swabs positive for *Pneumococcus* is recorded in families. A simple model might assume that the mean number of swabs is μ and the observation on the jth member of the ith family can be modelled by:

$$y_{ij} = \mu + \varepsilon_{ij}, \tag{12.33}$$

where ε_{ij} is a simple error term, with zero mean and variance σ^2, that is independent from observation to observation. However, this model does not reflect the fact that the observations are grouped within families. Moreover, if the variation between families is larger than that within families, then this cannot be modelled because only one variance has been specified. An obvious extension is to add an extra term, ξ_i, to (12.33) to accommodate variation between families. This term will be a random variable that is independent between families and of the ε_{ij}, has zero mean and variance σ_F^2. Thus, the new model for the jth member of the ith family is:

$$\mu + \xi_i + \varepsilon_{ij}.$$

It should be noted that it is the same realization of the random variable that is applied to each observation within a family; a consequence of this is that observations within a family are correlated. Two observations within a family have covariance σ_F^2 and, as each observation has variance $\sigma_F^2 + \sigma^2$, the correlation is

$$\sigma_F^2/(\sigma_F^2 + \sigma^2). \tag{12.34}$$

This is not surprising; the model is such that families with a propensity to exhibit pneumococcal infection will have a large value for ξ_i and as this is applied to each member of the family, each family member will tend to report a large value—that is, the values are correlated. Clearly, this tendency will be less marked if the within-family variation is substantial relative to that between families; this is reflected in (12.34) because, as σ^2/σ_F^2 becomes larger, (12.34) becomes smaller. It should be noted that correlations generated in this way cannot be negative: they are examples of the *intraclass correlation* discussed in §19.11.

Because they are random variables, the terms ξ and ε are referred to as *random effects* and their effect is measured by a variance or, more accurately, a *component of variance*, such as σ^2 and σ_F^2. More elaborate models can certainly be built. One possibility is to add extra terms that are not random (and so are often referred to as fixed effects) to elaborate on the simple mean μ. In Example 9.5 the families were classified into three categories measuring how crowded their living conditions were. The model could be extended to

$$\mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \xi_i + \varepsilon_{ij}, \tag{12.35}$$

where $x_{1i} = 1$ if the ith family lives in crowded conditions and is 0 otherwise, and $x_{2i} = 1$ if the ith family lives in uncrowded conditions and is 0 otherwise. The

parameter μ now measures the mean number of swabs positive for *Pneumococcus* in families living in overcrowded conditions. Note that the variables x_1 and x_2 need only a single subscript i because they measure quantities that only vary at the level of the family.

If the age of the family member was thought to affect the number of positive swabs then this could be incorporated into the model by allowing a suitable term in the model, such as

$$\mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3ij} + \xi_i + \varepsilon_{ij}, \tag{12.36}$$

where x_{3ij} is the age of the jth member of family i. As age varies between members of a family, the variable x_3 requires two subscripts: i, to indicate family, and j, to indicate the individual within the family. Of course, given the typical age differences within a family, the use of a linear term in this example is questionable but this can be overlooked for the purpose of illustration. Not only might the age of an individual affect the outcome but the rate of increase might vary between families. This can be incorporated by allowing the coefficient of age, β_3, to vary randomly between families. This can be built into the model by extending (12.36) to

$$\mu + \beta_1 x_{1i} + \beta_2 x_{2i} + (\beta_3 + \eta_i) x_{3ij} + \xi_i + \varepsilon_{ij}, \tag{12.37}$$

where β_3 is now the mean slope and η_i varies randomly between families with variance σ_b^2. The analyst can decide whether to insist that the random effects η_i and ξ_i are uncorrelated or to allow them to have covariance σ_{bF}. For the latter model the variance of the jth member of family i is

$$\sigma_b^2 x_{3ij}^2 + 2\sigma_{bF} x_{3ij} + \sigma_F^2 + \sigma^2. \tag{12.38}$$

It should be noted that allowing the slope to vary randomly has induced a variance that changes quadratically with age. Also, responses from members of the same family, say j and j', now have a correlation that depends on their ages, namely,

$$\frac{\sigma_b^2 x_{3ij} x_{3ij'} + \sigma_{bF}(x_{3ij} + x_{3ij'}) + \sigma_F^2}{\sqrt{[(\sigma_b^2 x_{3ij}^2 + 2\sigma_{bF} x_{3ij} + \sigma_F^2 + \sigma^2)(\sigma_b^2 x_{3ij'}^2 + 2\sigma_{bF} x_{3ij'} + \sigma_F^2 + \sigma^2)]}}.$$

In this way a family of models can be defined which allow many forms of data to be analysed.

Method of Estimation

For some purposes and sufficiently regular problems, apparently *ad hoc* methods are optimal. For example, suppose the aim is to compare the mean number of swabs positive for *Pneumococcus* between crowded and overcrowded families

in Example 9.5. A simple analysis would be to compute the mean number of swabs in each family and compare the two groups, using the six family means in each group as the outcome variable. If model (12.35) obtained, then the difference in group means would estimate β_1 and the pooled within-group variance would estimate $\sigma_F^2 + n^{-1}\sigma^2$, where n is the (constant) size of each family. It is perhaps not surprising that no new methodology is needed for this analysis, as the split-unit analysis of variance described in §9.6 can provide a complete analysis of these data.

The split-unit analysis of variance could still cope if the number of families at each level of crowding were unequal, but the method would fail if the number of people within each family were not constant. The mean for the ith family would have variance $\sigma_F^2 + n_i^{-1}\sigma^2$, where n_i is the size of family i. As this varies between families, an unweighted mean of the families would not necessarily be the optimal way to compare levels of crowding. However, the optimal weighting will depend on the unknown value of the ratio σ_F^2/σ^2; in general, the optimal weighting will depend on several parameters whose values will have to be estimated. A satisfactory approach to the general problem of analysing hierarchical data requires methodology that can handle this kind of problem. A more sophisticated problem is that the analysis should not only be able to estimate the parameters that determine the appropriate weights, but should allow estimates of error to be obtained that acknowledge the uncertainty in the estimates of the weights.

Suppose the 90 observations in Example 9.5 are written as a 90×1 vector y, then the model in (12.35) can be written:

$$y = X\beta + \delta, \tag{12.39}$$

where δ is a 90×1 vector of error terms that subsume the terms ξ and ε from (12.35), X is a 90×2 matrix and β is a 2×1 vector. Consequently δ has zero mean and dispersion matrix V. The form of V is determined by the structure of the random effects in the model and will be specified in terms of the variance parameters. In this example, V has the form:

$$V = \begin{pmatrix} V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_{18} \end{pmatrix}, \tag{12.40}$$

i.e. V has a block-diagonal structure where the only non-zero elements are those in the submatrices, V_i, shown. The matrix V_i is the dispersion matrix of the observations from family i. In general, this could be an $n_i \times n_i$ matrix but, as all the families in this example are of size 5, each V_i is a 5×5 matrix. As has been noted, the variance of each response is $\sigma_F^2 + \sigma^2$ and the covariance between any two members of the same family is σ_F^2, so each V_i is

$$\begin{pmatrix} \sigma_F^2+\sigma^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 \\ \sigma_F^2 & \sigma_F^2+\sigma^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 \\ \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma^2 & \sigma_F^2 & \sigma_F^2 \\ \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma^2 & \sigma_F^2 \\ \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2 & \sigma_F^2+\sigma^2 \end{pmatrix}.$$

If the values of σ^2, σ_F^2 were known, then the estimator of the β parameters in (12.39) having minimum variance would be the usual generalized least squares estimator (see (11.54)):

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \tag{12.41}$$

As σ^2, σ_F^2 are unknown, the estimation proceeds iteratively. The first estimates of β are usually obtained using ordinary least squares, that is assuming V is the identity matrix. An estimate of δ can then be obtained as $\hat{\delta} = y - X\hat{\beta}$. The 90×90 matrix $\hat{\delta}\hat{\delta}^T$ has expectation V and the elements of both these matrices can be written out as vectors, simply by stacking the columns of the matrices on top of one another. Suppose the vectors obtained in this way are W and Z, respectively, then Z can in turn be written $\sum \sigma_k^2 z_k$, where the z_i are vectors of known constants. In the case of Example 9.6, $\sigma_1^2 = \sigma_F^2, \sigma_2^2 = \sigma^2$ and the z vectors comprise 0s and 1s. A second linear model can now be fitted using generalized least squares with W as the response, the *design matrix* comprising the z vectors and the parameter estimates being the estimates of the variance components defining the random effects in the model; further details can be found in Appendix 2.1 of Goldstein (1995). New estimates of the βs can be obtained from (12.41), with V now determined by the new estimates of the variance components. The whole process can then be repeated until there is little change in successive parameter estimates. This is essentially the process used by the program MLwiN (Goldstein *et al.*, 1998) and is referred to as *iterative generalized least squares* (IGLS).

If the approach outlined above is followed exactly, then the resulting estimates of the variance components will be biased downwards. This is because in the part of the algorithm that estimates the random effects the method uses estimates of fixed effects as if they were the correct values and takes no account of their associated uncertainty. This is essentially the same problem that arises because a standard deviation must be estimated by computing deviations about the sample mean rather than the population mean. In that case, the solution is to use $n - 1$ in the denominator rather than n. A similar solution, often referred to as restricted maximum likelihood (see Patterson & Thompson, 1971), can be applied in more general circumstances, such as those encountered in multilevel models, and is then called *restricted iterative generalized least squares* (RIGLS).

A complementary problem arises from neglecting uncertainty in estimates of the random effects. Standard theory allows values for the standard errors of the

parameter estimates to be obtained; for example, the dispersion matrix of the estimates of the fixed parameters can be found as $(X^{\mathrm{T}}V^{-1}X)^{-1}$. In practice, V is evaluated using the estimated values of the variance components but the foregoing formula takes no account of the uncertainty in these estimates and is therefore likely to underestimate the required standard errors. For large data sets this is unlikely to be a major problem but it could be troublesome for small data sets. A solution is to put the whole estimation procedure in a Bayesian framework and use diffuse priors. Estimation using Markov chain Monte Carlo (MCMC) methods will then provide estimates of error that take account of the uncertainty in the parameter estimates. For a fuller discussion of the application of MCMC methods to multilevel models, see Appendix 2.4 of Goldstein (1995) and Goldstein *et al.* (1998). The use of MCMC methods in Bayesian methodology is discussed in §16.4.

More generally, this matter does, of course, raise the question of what constitutes a small or large data set, as this is not entirely straightforward when dealing with hierarchical data. There is no longer a single measure of the size of a data set; the implications of having 400 observations arising from a measurement on each of 20 patients in each of 20 general practices will be quite different from those arising from measuring 100 patients in each of four practices. Broadly speaking, it is important to have adequate replication at the highest levels of the hierarchy. If the model in (12.35) were applied to the example of data from general practices, with ξ representing practice effects, only one realization of ξ would be observed from each practice, and a good estimate of σ_F^2 therefore requires that an adequate number of practices be observed. Attempts to try to compensate for using an inadequate number of practices by observing more patients in each practice will, in general, be futile.

Estimation of residuals

In §11.9 simple regression models were checked by computing residuals. Residuals also play an important role in the more elaborate circumstances of multilevel models, and indeed have more diverse uses.

The residuals are clearly useful for checking the assumptions of the model; for example, normal probability plots of the residual effects at each level allow the assumptions underlying the random effects within the model to be assessed. It should, however, be noted that there may be more than one set of residuals at a given level, since there will be a separate set corresponding to each random effect. For example, in (12.37) there will be a set of residuals at the level of the individual, corresponding to ε_{ij}, but there will also be two sets of residuals at the level of the family, corresponding to η_i and ξ_i.

However, in addition to their role in model checking, the residuals can be thought of as estimates of the realized values of random effects. This can be

potentially useful, especially for residuals at levels above the lowest; for example, in a study of patients in general practices the practice-level residual might be used to help place the specific practice in the context of other practices, once the influence of other effects in the model had been taken into account. In particular, they might be used in attempts to rank practices. However, attempts to rank units in this way are fraught with difficulty and should be undertaken only with great circumspection; see Goldstein and Spiegelhalter (1996) for a fuller discussion.

The extension of the idea of residuals beyond those for a standard multiple regression (see §11.9) gives rise to complexities in both their estimation and their definition. There is considerable merit in viewing the process as *estimating random effects* rather than as an exercise in extending the definition of a residual in non-hierarchical models. Indeed, there is much relevant background material in the article by Robinson (1991) on estimating random effects.

As a brief and incomplete illustration of the issues, consider model (12.35). The family-level residuals are $\{\hat{\xi}_i\}$ and these are 'estimates' of the random variables $\{\xi_i\}$ that appear in the model. As Robinson (1991) discusses, some statisticians are uneasy about this, seeing it as lying outside the realm of parameter estimation. However, even if such objections are accepted, there is likely to be little objection to the notion of predicting a random effect and the usual definition for residuals in a multilevel model is often put in this way, namely,

$$\hat{\xi}_i = \mathrm{E}(\xi_i | y, \hat{\mu}, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_F^2, \hat{\sigma}^2).$$

If the 'raw' residuals are defined as $r_{ij} = y_{ij} - \hat{\mu} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$ and the mean raw residual for family i is $\bar{r}_i = \sum_{j=1}^{n_i} r_{ij}/n_i$ then the above expectation can be expressed as:

$$\hat{\xi}_i = \frac{\hat{\sigma}_F^2}{\hat{\sigma}_F^2 + \hat{\sigma}^2/n_i} \bar{r}_i. \tag{12.42}$$

It might have been expected that the family-level residual would simply be $\bar{r}_i$ but (12.42) is used instead. The difference, namely the factor $\hat{\sigma}_F^2/(\hat{\sigma}_F^2 + \hat{\sigma}^2/n_i)$, is often referred to as a shrinkage factor, as its effect is to 'shrink' $\bar{r}_i$ towards zero because the factor is always between 0 and 1. If the within-family variation is small compared with that between families, or if the size of the family is large, the shrinkage is minor. However, for small families the effect can be noticeable. The reasons for the appearance of this factor can best be appreciated if the procedure is considered in the context of estimation. Information on the term ξ_i is obtained from observation on family i. If substantial information is available from the family, then the estimate $\bar{r}_i$ is essentially sound. However, if few observations are available within a given family the method provides an estimate that is a compromise between the observed values and the population mean of the ξs,

namely, 0. It should be noted that this definition naturally leads to individual-level residuals being defined to be $r_{ij} - \hat{\xi}_i$ rather than $r_{ij} - \bar{r}_i$.

If the residuals at levels above the lowest are to be used in their own right, perhaps in a ranking exercise for higher-level units, it may be necessary to compute appropriate standard errors and interval estimates. For discussion of these issues, the reader should consult Appendix 2.2 in Goldstein (1995).

Example 12.6

A trial was conducted to assess the benefit of two methods of giving care to patients who had recently been diagnosed as having Type II diabetes mellitus. The trial was run in general practices in the Wessex region of southern England with 250 patients from 41 practices, 21 randomized to the group in which nurses and/or doctors in the practice received additional training on patient-centred care (the *intervention* group) and patients in the other 20 practices received routine care (the *comparison* group). As the data comprise patients within practices, it is appropriate to use a method of analysis for hierarchical data and a multilevel model is used for the analysis of data on body-mass index (BMI: weight of patient over the square of their height, in kg/m^2). Several outcomes were measured and further details can be found in Kinmonth *et al.* (1998). In that report simpler methods were used because, as will be demonstrated, the variation between practices is not substantial compared with that within a practice.

The modelling approach reported here is based on a subset of 37 practices and 220 patients. The model fitted has four fixed effects and two random effects. The four fixed effects are a general mean and three binary variables: (i) a variable indicating the treatment group to which the practice was allocated in the randomization, $x_1 = 0$ for comparison and $x_1 = 1$ for the intervention group; (ii) a variable indicating whether the number of patients registered with the practice was above 10 000, $x_2 = 0$, or below, $x_2 = 1$; and (iii) a variable indicating whether care was always given to these patients by a nurse, $x_3 = 0$, or otherwise, $x_3 = 1$. There are random effects for the practices, with variance σ_P^2, and for the patients, with variance σ^2, so the full model is

$$y_{ij} = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \xi_i + \varepsilon_{ij}.$$

The term for a treatment effect obviously must be present in the model and the terms for the size of the practice and the care arrangements within a practice are included because these were used to stratify the allocation procedure. In this instance, no fixed effects vary at the patient level. If the model is fitted using RIGLS, the estimates of the parameters are as follows (fixed effects in kg/m^2, variances in kg^2/m^4):

Parameter	μ	β_1	β_2	β_3	σ_P^2	σ^2
Estimate	28.67	1·69	0·11	0·90	0·99	34·85
SE	1·15	0·90	0·95	1·18	1·57	3·58

The mean BMI is estimated to be 1·7 kg/m^2 higher in the intervention group than in the comparison group.

The estimate of σ_P^2 is considerably smaller than its standard error, which is the basis for noting that an analysis which ignores the clustering of patients into practices is unlikely to be misleading. Confidence intervals for the parameters, in particular the treatment effect, can be constructed in the usual way provided that the random effects can be assumed to follow normal distributions. This is checked in Fig. 12.10; note that panel (a) contains 220 points, as it corresponds to the random effects for patients, while panel (b), which corresponds to the practice random effect, has only 37 points. The plots are certainly reasonable confirmation of the assumption of normality, although the pattern for larger residuals at the patient level suggests that some further analysis may be helpful. Note the different scales, which reflect the marked difference in the sizes of the estimates of σ^2 and σ_P^2.

The estimates given above are from a method, RIGLS, that takes account of the uncertainty in the fixed effects when these are used to find estimates of random effects, but the quoted standard errors take no account of the uncertainty in the estimates of random effects. To do this a method based on a Bayesian approach, with diffuse priors and estimation using a Gibbs sampler (see §16.4), would be required. The estimate of treatment effect, its standard error and a 95% confidence interval were computed for each of three methods of estimation. The first is IGLS, the second is RIGLS and the last is a Bayesian formulation with fixed effects having normal prior distributions, with very large variance, and random effects having priors that are uniform between 0 and a very large value.

Method	Treatment effect	Standard error	95% confidence interval estimate
IGLS	1·64 kg/m^2	0·85 kg/m^2	(−0·03, 3·31) kg/m^2
RIGLS	1·69 kg/m^2	0·90 kg/m^2	(−0·07, 3·45) kg/m^2
Gibbs sampler	1·76 kg/m^2	1·01 kg/m^2	(−0·19, 3·79) kg/m^2

The point estimates of treatment effect are very similar and, for all practical purposes, identical. The Gibbs sampler is based on a chain of length 40 000, and technical diagnostic values suggest that this is adequate to ensure convergence of the method. In this case, the 95% confidence interval estimate is derived from the distribution of estimates of treatment effect found from the chain. It is also clear that as the estimation method takes into account sources of variation neglected in other approaches, the estimate of standard error, and hence the width of the interval estimate, increases.

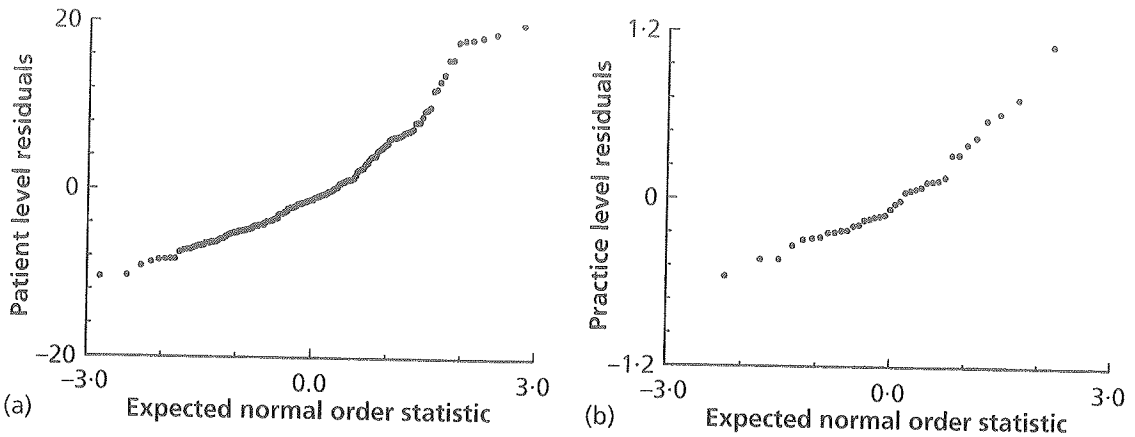

Fig. 12.10 Normal probability plots for the estimated residuals at the patient and practice levels from the trial of patient-centred care in general practice.

All the interval estimates demonstrate that any advantage the intervention group might have over the comparison group is practically negligible, whereas the comparison group could be substantially better than the intervention group.

Other uses of multilevel models

It has already been mentioned that multilevel models can be very useful in the analysis of hierarchical data when the response is not continuous, but even with continuous responses there is still scope for useful extensions. A few of these are outlined below.

Variations in random effects

In model (12.37) the variance increases with the square of age according to (12.38). However, it may be that this is inappropriate and a linear dependence is required. In this case the same model can be fitted, but in the fitting procedure the parameter σ_b^2 is held at 0. This is perhaps best regarded as a device for fitting the appropriate model, as the notion of a variable with a non-zero covariance but zero variance is not easy to interpret.

It should also be pointed out that the variation at the lowest level of the hierarchy, which hitherto has been assumed to be constant, can be made more elaborate. For example, allowing different variances at the lowest level of the hierarchy for different groups is possible.

As will be seen in the next section, observations made longitudinally on, for example, a patient are often serially correlated. A reasonable supposition is that the correlation is larger between measurements made closer together in time than further apart. However, the correlations induced between measurements on the same patient by a model which is of the form (12.34) are the same regardless of their separation in time. A model which overcomes this is known as the *autocorrelation* or *autoregressive model*, and a simple version of this leads to a correlation of $\rho^{|r-s|}$ between observations at times r and s. This kind of feature can be accommodated in multilevel models; see Goldstein *et al.* (1994) for details and §12.7 for a further discussion of autoregressive processes.

Non-linear models

In all the multilevel models discussed hitherto the response has depended linearly on the covariates. Non-hierarchical non-linear models were discussed in §12.4 and these can be extended to hierarchical data. Such models can be particularly useful for growth data; the data are hierarchical because individuals are measured longitudinally, but adequate modelling of the form of the response usually requires non-linear functions.

Non-linear models can be accommodated by repeatedly using Taylor expansions to linearize the model. There are close connections between this way of extending linear multilevel models and the types of model obtained by extending non-linear models to hierarchical data. Models generated in this way have recently been widely used to model pharmacokinetic data and this alternative approach is well described by Davidian and Giltinan (1995).

Multivariate analysis

Multivariate analysis, which is considered at greater length in Chapter 13, is the term used to describe a collection of statistical techniques which can be used when each observation comprises several variables, that is, each observation is a vector of, say, p dimensions. For example, the concentrations of creatinine, sodium and albumin in the blood of a patient may be measured, yielding as an observation a three-dimensional vector, which will have a mean that is also a three-dimensional vector and its 'variance' is a 3×3 dispersion matrix of variances and covariances. Of course, each component of this vector will be on its own scale and it will not, in general, be possible to specify common parameters across components of the vector.

Deployment of a certain amount of ingenuity means that multivariate observations of this kind can be analysed as multilevel models, and it turns out that there are some attractive benefits to viewing such data in this way. Suppose that $\boldsymbol{y}_i$ is a p-dimensional vector observed on patient i. It is assumed that if $\boldsymbol{y}$ were a scalar then this would be the lowest level of a hierarchy but, despite the single subscript, there is no implication that the patient is the top level of a hierarchy; the subscript might, for example, be elaborated to describe a hierarchy in which the patient is from a hospital, which in turn belongs to a given health authority. The multivariate nature of $\boldsymbol{y}$ is accommodated by the device of constructing a new lowest level to the hierarchy, which describes the variables within each vector $\boldsymbol{y}_i$. This makes extensive use of dummy, i.e. 0–1, variables. To be specific, suppose the scalar y_{ij} is the jth variable observed on patient i (so, for example, in the above instance, y_{i1} might be the creatinine, y_{i2} the sodium and y_{i3} the albumin on patient i); the model used is then

$$y_{ij} = \beta_1 z_{1ij} + \beta_2 z_{2ij} + \beta_3 z_{3ij} + \xi_{1i} z_{1ij} + \xi_{2i} z_{2ij} + \xi_{3i} z_{3ij},$$

where z_{1ij} is 1 if y_{ij} is an observation on creatinine (i.e. the first variable in the vector $\boldsymbol{y}_i$) and 0 otherwise. Similarly z_{kij} is 1 or 0 and is 1 only if y_{ij} is an observation on the kth variable in the vector. The random effects are at the level one up from the lowest level, i.e. the patient level in this example, and, in general, have arbitrary variances and covariances. Note that there are no random effects at the lowest level as this level is simply a device to distinguish between the different variables within each observed vector.

A notable advantage to this way of specifying multivariate data is that there is no requirement that each variable is present in all the vectors—that is, the vector can be only partially observed for some patients. If, for example, the albumin is not measured on patient i then the model simply has no entry for y_{i3}. This can be very useful because incomplete vectors can cause serious difficulties for standard approaches to multivariate analysis. It can be helpful if an element of a vector is inadvertently missing, although the analyst must then be satisfied that the omission is not for a reason that could bias the analysis (related concerns to this arise in the analysis of longitudinal data and are discussed at more length in the next section and also in clinical trials (see §18.6)). It can also be useful to arrange to collect data in a way that deliberately leads to the partial observation of some or all vectors. If the creatinine, sodium and albumin of the foregoing example are to be observed on premature infants then it may not be permissible to take sufficient blood for all items to be observed on each baby. It may then be helpful to observe just two of the variables on each infant, to arrange the patients into three groups and take just one of the three possible pairs of measurements on each baby. This is a simple example of a *rotation design*, in which predetermined subgroups of the variables of interest are observed on particular individuals. Further details on rotation designs and on the application of multilevel models to multivariate methods can be found in Goldstein (1995, Chapter 4).

12.6 Longitudinal data

In many medical studies there is interest not only in observing a variable at a given instant but in seeing how it changes over time. This could be because the investigator wishes to observe how a variable evolves over time, such as the height of a growing child, or to observe the natural variation that occurs in a clinical measurement, such as the blood pressure of a volunteer on successive days. A very common reason is to observe the time course of some intervention, such as a treatment: for example the respiratory function of a patient at a series of times after the administration of a bronchodilator, or the blood glucose of a diabetic patient in the two hours after a glucose challenge.

Data collected successively on the each of several units, whether patients, volunteers, animals or other units, are variously referred to as *longitudinal data, serial data* or *repeated measurements*, although many other terms are encountered from time to time. Typically the data will be collected on several, say, g, groups of individuals, perhaps defined by allocation to different treatments; typically there will be n_i units in the ith group. The jth unit in the ith group will be observed k_{ij} times. There is wide variation between studies in the timing and number of the observations on an individual. The observations on each individual constitute a *time series*, and in the next section methods that are

traditionally described as applying to time series are discussed. However, such methods apply to a single long series, perhaps comprising hundreds of observations, whereas the data discussed in this section typically arise from many shorter series, often of two to 20 measurements per individual.

Another feature that varies widely between studies is why observations are made when they are. Most studies attempt to make observations at preplanned times; those which do not—for example, taking observations opportunistically, or perhaps when some clinical event occurs—are likely to present formidable problems of interpretation. For preplanned observations there is no requirement that they be taken at regular intervals, and in fact it may often not be sensible to do so; for example, observations may need to be taken more frequently when the response is changing rapidly, provided, of course, that that aspect of the response is of interest. For example, in the study of the profile of the blood level of a short-acting drug, measurements may be made every 10 or 15 min in the initial stages when the profile is changing rapidly, but then less frequently, perhaps at 1, 2 and 3 h post-administration. In many studies in the medical literature the reasons behind the timing of observations are seldom discussed, and it may be that this aspect of research involving the collection of longitudinal data would benefit from greater reflection.

Often it will be intended to measure individuals at the same set of times, but this is not achieved in every case. Such *missing data* give rise to two separate problems, which are often not distinguished from one another as clearly as they might be. The first, which is largely technical, is that the varying number of observations per individual may influence the type of analysis performed, as some methods are less tractable, or even impossible, when the number of observations varies between individuals. The second problem, which is more subtle and, because it can evade an unwary analyst, potentially more serious, is that the missing data are absent for reasons related to the purpose of the study, so an analysis of only the available data may well be biased and possibly misleading. The second of these problems will be discussed at greater length towards the end of the present section.

As with any statistical analysis, it is important when dealing with longitudinal data that their structure is respected. The two most important aspects for longitudinal data are: (i) that the method should take account of the link between successive measurements on a given individual; and (ii) that it should recognize that successive measurements on an individual will not, in general, be independent. Despite warnings to the contrary (Matthews *et al.*, 1990; Matthews, 1998), both these aspects appear to be frequently overlooked in the medical literature, where it is common to see separate analyses performed at the different times when measurements were made. Such analyses ignore the fact that the same individuals are present in successive analyses and makes no allowance for within-individual correlation.

Appropriate methods for the analysis of longitudinal data have been studied intensively in recent years and there is now a very large statistical literature on the subject. Some topics will be discussed in more detail below, but the reader should be aware that it is a highly selective account. The selection has largely been guided by consideration of the issues outlined in the previous paragraph, and in particular focuses on methods available for studying correlated responses. Important areas, such as growth curves, which are concerned with the shape of the response over time, are not mentioned and the reader should consult one of the excellent specialist texts in the field, such as Crowder and Hand (1990) or Diggle *et al.* (1994). Methods for graphing longitudinal data are also not covered. This is a surprisingly awkward but practically important problem, which has received much less attention than analytical methods; articles which contain some relevant material include Jones and Rice (1992), and Goldstein and Healy (1995), as does Chapter 3 of Diggle *et al.* (1994).

Repeated measures analysis of variance

As was remarked in §12.5, the multiple measurements taken on a patient means that longitudinal data can be viewed as a form of hierarchical data. It was also noted in the previous section that a split-unit analysis of variance (see §9.6) could analyse certain forms of sufficiently regular hierarchical data. It follows that split-unit analysis of variance can be pressed into service to analyse longitudinal data, with the whole units corresponding to the individuals and the subunits corresponding to measurement occasion. When used in this context the technique is usually referred to as *repeated measures analysis of variance*.

This method requires that each individual be measured on the same number of occasions, say, k times, that is, $k_{ij} = k$. There is no requirement that the number of individuals in each group is the same. If the total number of individuals in the study is denoted by $N = \sum_{i=1}^{g} n_i$, the analysis of variance table breaks down as follows into two strata, one between individuals and one within individuals.

Source of variation	DF	
Between-individuals stratum	$N-1$	
Groups		$g-1$
Residual between individuals		$N-g$
Within-individuals stratum	$N(k-1)$	
Occasions		$k-1$
Occasions × groups		$(g-1)(k-1)$
Residual within individuals		$(N-g)(k-1)$
Grand total	$Nk-1$	

Use of this technique therefore allows the analyst to assess not only effects of time (the Occasions row) and differences between groups (the Groups row), but whether or not the difference between groups changes with time (the Occasions × groups interaction). This is often the row of most interest in this technique. However, some care is needed; for example, if the groups arise through random allocation of individuals to treatments and the first occasion is a pretreatment baseline measurement, then any treatment effect, even if it is constant across all times post-randomization, will give rise to an interaction because there will necessarily be no difference between the groups on the first occasion. A minor problem is that, if there is a significant interaction between occasions and groups, then it is natural to ask when differences occur. If there is a prior belief of a particular pattern in the response, then this can be used to guide further hypothesis tests. In the absence of such expectations, techniques that control the Type I error rate need to be considered; the discussion in §8.4 is relevant here. However, in applying these methods, it is important that the user remembers the ordering implicit in the Occasions term and the fact that, in general, the response is likely to change smoothly rather than abruptly with time.

A further problem with the method is that the variance ratios formed in the within-individuals stratum will not, in general, follow an F distribution. This is a consequence of the dependence between successive measurements on the same individual. The use of an F test is valid under certain special circumstances but these are unlikely to hold in practice. Adjustments to the analysis can be made to attempt to accommodate the dependence under other circumstances and this can go some way to salvaging the technique. The adjustments amount to applying the usual method but with the degrees of freedom for the hypothesis tests for occasions and occasions × groups reduced by an appropriate factor. More details of this are given below, but a heuristic explanation of why this is a suitable approach is because within-individual correlation means that a value on an individual will contain some information about the other values, so there are fewer independent pieces of information than a conventional counting of degrees of freedom would lead you to believe. It is therefore sensible to apply tests with a reduced number of degrees of freedom.

In order to be more specific, suppose that $\boldsymbol{y}_i$ is the k-dimensional vector of observations on individual i and the dispersion matrix of this vector is $\boldsymbol{\Sigma}$. The between-individuals stratum of the repeated measures analysis of variance can be viewed as a simple one-way analysis of variance between groups on suitably scaled individual totals, namely, the values proportional to $\mathbf{1}^{\mathrm{T}}\boldsymbol{y}_i$, where $\mathbf{1}$ is a k-dimensional vector of ones. The within-individual stratum is a simultaneous analysis of the within-individual contrasts, namely, of the $\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{y}_i$, where $\boldsymbol{a}_1, \ldots \boldsymbol{a}_{k-1}$ are $k-1$ independent vectors each of whose entries sum to zero, i.e. $\boldsymbol{a}_j^{\mathrm{T}}\mathbf{1} = 0$. The variance ratios in this analysis will be valid if the dispersion matrix is such that $\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_j$ is proportional to the identity matrix (see §11.6). One

form for Σ that satisfies this is the *equi-correlation structure*, where all variances in Σ are equal, as are all the covariances. However, this implies that pairs of observations taken close together in time have the same correlation as pairs taken at widely separated times, and this is unlikely to hold in practice.

By using the work of Box (1954a, b), Greenhouse and Geisser (1959) devised an adjustment factor that allows the technique to be applied for an arbitrary dispersion matrix. The adjustment is to reduce the degrees of freedom in the hypothesis tests for occasions and for occasions × groups by a factor ε, so, for example, the test for an effect of occasions compares the usual variance ratio statistic with an F distribution on $(k-1)\ \varepsilon$ and $(N-g)(k-1)\ \varepsilon$ degrees of freedom. The factor ε, often called the Greenhouse–Geisser ε, is defined as

$$\varepsilon = \frac{\{\mathrm{tr}(\boldsymbol{\Sigma H})\}^2}{(k-1)\mathrm{tr}(\boldsymbol{\Sigma H \Sigma H})}, \tag{12.43}$$

where $\boldsymbol{H} = \boldsymbol{I}_k - \frac{1}{k}\boldsymbol{J}_k$, where $\boldsymbol{I}_k$ is a $k \times k$ identity matrix and $\boldsymbol{J}_k$ is a $k \times k$ matrix of ones. If this correction is applied, then the variance ratios in the within-patient stratum still do not follow an F distribution but the discrepancy is less than would be the case without the correction. Of course, in practice ε must be estimated by substituting an estimate of Σ into (12.43) and the properties of the test based on an estimated ε may not be the same as those using the true value. Huynh and Feldt (1976) devised an alternative correction of similar type whose sampling properties might be preferable to those of (12.43).

Calculation of (12.43) is awkward, requires an estimate of Σ and may have uncertain properties when ε has to be estimated. A practically useful device is available because it can be shown that (12.43) must lie between $(k-1)^{-1}$ and 1. As the degrees of freedom decrease, any critical point, say the 5% point, of the F distribution will increase. So, if an effect is not significant in an uncorrected test in the within-individual stratum, then it will not become significant when the correction is applied. Similarly, if an effect is significant when a correction using $(k-1)^{-1}$ is used rather than ε, then it would be significant at that level if the correction in (12.43) were used. Using this approach the analyst only has to compute an estimate of ε for effects which are significant under the uncorrected analysis but not under the analysis using the factor $(k-1)^{-1}$.

The between-individuals stratum is also not without problems. The test for equality of group means, which would only sensibly be considered in the absence of an occasion by group interaction, is generally valid (given usual assumptions about normality and equality of variances) because it is essentially the one-way analysis of variance of the means of the responses on an individual. This amounts to summarizing the response of an individual by the mean response, and this is an automatic consequence of using repeated measures analysis of variance. However, the mean response over time may not be an appropriate way to measure a relevant feature of the response. The idea of reducing the k

responses on an individual to a suitable scalar quantity, which can then be analysed simply, is the key idea behind the important approach to the analysis of longitudinal data that is outlined in the next subsection.

Summary measures

Perhaps the principal difficulty in analysing longitudinal data is coping with the dependency that is likely to exist between responses on the same individual. However, there is no more difficulty in assuming that responses from different individuals are independent than in other areas of data analysis (this assumes, for simplicity, that the individuals in the analysis are not embedded in a larger design, such as a complex survey (see §19.2) or a cluster-randomized trial (see §18.9), which may itself induce dependence). Consequently, if the responses on individual i, $\boldsymbol{y}_i$, together with other information, such as the times of the responses, $\boldsymbol{t}_i$, say, are used to compute a suitable scalar (i.e. a single value), s_i, say, then the s_i are independent and can be analysed using straightforward statistical methods. The value of this approach, which can be called the *summary measures method*, rests on the ability of the analyst to be able to specify a suitable function of the observations that can capture an important feature of the response of each individual. For this reason the method is sometimes referred to as *response feature analysis* (Crowder & Hand, 1990). The method has a long history, an early use being by Wishart (1938). More recent discussions can be found in Healy (1981), Yates (1982) and Matthews *et al.* (1990).

If, for example, the response of interest is the overall level of a blood chemistry measurement, then the simple average of the responses on an individual may be adequate. If the effect of some treatment on this quantity is being assessed then it may be sensible to omit the first few determinations from the average, so as to allow the treatment time to have its effect. A rate of change might best be summarized by defining s_i to be the regression slope of $\boldsymbol{y}_i$ on $\boldsymbol{t}_i$. Summaries based on the time-scale may be particularly important from a clinical point of view: the time that a quantity, such as a drug concentration, is above a therapeutic level or the time to a maximum response may be suitable summaries. It may be that more than one feature of the data is of interest and it would then be legitimate to define and analyse a summary for each such feature. Simple bivariate analyses of summaries are also possible, although they seem to be little used in practice. Judgement should be exercised in the number of summaries that are to be analysed; summaries should correspond to distinct features of the response and in practice there are unlikely to be more than two or three of these.

The choice of summary measure should be guided by what is clinically or biologically reasonable and germane to the purpose of the study. Indeed, it is preferable to define the summary before the data are collected, as this may help to focus attention on the purpose of the investigation and the most appropriate

times at which to make observations. This is particularly important when time-based summaries, such as time above a therapeutic level, are considered. Any prior information on when concentrations are likely to reach and decline from therapeutic levels can lead the investigators to placing more observations around these times. Occasionally, theoretical background can inform a choice of summary measure; the maximum response and the area under the response versus time curve are summaries that have long been used in pharmacology for such reasons. Choice of summary on the basis of the observed responses can be useful but, unless the summary ultimately chosen has a clear biological or clinical interpretation, the value of this approach is much reduced. Healy (1981) outlines the role of orthogonal polynomials in the method, although this is probably of greater theoretical importance in relating the method to other approaches than of practical interest.

There are, of course, drawbacks to the method. The most obvious is that in some circumstances it may not be possible to define a summary that adequately captures the response over time. Other problems in longitudinal data analysis are not naturally approached by this method; assessing whether changes in the blood concentration of a beta-blocker are related to changes in blood pressure is an example. Also there are technical problems. Many of the simple statistical methods that it is assumed will be used for the analysis of the summaries suppose that they share a common distribution, except perhaps for a few parameters such as those describing differences in the mean between treatment groups. In particular, the variances will often be assumed equal. This can easily be violated and this is illustrated by the following situation. Suppose the summary chosen is the mean and that the model for the elements of y_i is

$$y_{ij} = \mu_i + \xi_i + \varepsilon_{ij}, \tag{12.44}$$

where ξ_i and ε_{ij} are random variables with zero mean and $\text{var}(\xi_i) = \sigma_B^2$ and $\text{var}(\varepsilon_{ij}) = \sigma^2$. The mean of the elements of y_i has variance $\sigma_B^2 + n_i^{-1}\sigma^2$ and this will differ between individuals unless all the n_i are equal. While the intention at the outset of the study may have been to ensure that all the n_i were equal, it is almost inevitable that some individuals will be observed incompletely. However, even if there are marked differences between the n_i, there will be little important difference in the variances if the between-individuals variance, σ_B^2, is substantial relative to the within-individual variance, σ^2. Obviously there may be circumstances when concerns about distributional aspects of summary measures may be less easy to dismiss.

The problem with unequal variance for the mean response arose from unequal replication, which can commonly be attributed to occasional missing data. This leads naturally to the problem of dealing with missing values, which are of concern throughout statistics but seem to arise with especial force and frequency in longitudinal studies. On a naïve level the method of summary

measures is sufficiently flexible to deal with missing data; a summary such as a mean or regression slope can often be computed from the observations that are available. However, to do so ignores consideration of *why* the unavailable observations are missing and this can lead to a biased analysis. This is clearly illustrated if the summary is the maximum observed response: if the response is measured at weekly visits to an out-patient clinic and large values are associated with feeling especially unwell, then it is precisely when the values of most interest arise that the patient may not feel fit to attend for observation. The problem of missing data, which is discussed in more detail at the end of this section, is only a special problem for the method of summary measures in so far as the method may make it too easy to overlook the issue altogether, and this should not be a problem for the alert analyst.

Modelling the covariance

A natural method for dealing with longitudinal data is to view the response on an individual as a vector from a suitable multivariate distribution, typically a multivariate normal distribution. In this way the dependence is handled by assuming each vector has a dispersion matrix Σ; if each vector has k elements then the $\frac{1}{2}k(k+1)$ parameters describing the dispersion are estimated from the data in the usual way for multivariate analysis. For example, two groups could be compared using Hotelling's T^2 statistic (see Mardia *et al.*, 1979, pp. 139 ff.). A good discussion of the application of multivariate methods to longitudinal data can be found in Morrison (1976).

There are, however, good reasons why this approach is seldom adopted. As with many medical applications of multivariate methods (see multiple outcomes in clinical trials, §18.3), these general methods are rather inefficient for specialized application. In the case of longitudinal data analysis the dispersion matrix may plausibly take forms in which correlations between occasions closer in time are higher, rather than the general form allowed by this class of methods. Differences in mean vectors might also be expected to change smoothly over time. In addition, the ever-present problem of missing values means that in practice not all vectors will be of the same length and this can cause substantial problems for standard multivariate methods.

An alternative way to view substantially the same analysis, but which readily accommodates unequal replication within each individual, is to put the analysis in terms of a linear model. If the vectors y_i from the N individuals in the study are stacked to form a single M-dimensional vector y (where M is the number of observations in the study), then this can be written with some generality as $y = X\beta + \epsilon$, where X is an $M \times p$-dimensional design matrix and β is a p-dimensional vector describing the mean response. The longitudinal nature of the data is described by the form of the dispersion matrix of ϵ, namely Σ. This is

block diagonal, as in (12.40), now with N blocks on the diagonal, and the dispersion matrix for the ith individual in the study, Σ_i, is the ith block. Usually the Σ_i are different because of missing data, so each Σ_i comprises the available rows and columns from a common 'complete case' matrix Σ_c.

General statistical theory tells us that the best estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} X_i^{\mathrm{T}} \Sigma_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^{\mathrm{T}} \Sigma_i^{-1} y_i\right), \tag{12.45}$$

where X_i is the matrix comprising the rows of X that relate to individual i. However, this estimator is only available if the Σ_i are known and this will hardly ever be the case. An obvious step is to use (12.45) with an estimate, $\hat{\Sigma}_i$, in place of Σ_i. If all Σ_i were the same, such an estimator would be:

$$\frac{1}{N-p} \sum_{i=1}^{N} (y_i - X_i \hat{\boldsymbol{\beta}})(y_i - X_i \hat{\boldsymbol{\beta}})^{\mathrm{T}}.$$

In the case of missing data there is no simple solution; a sensible approach is to estimate the (u, v)th element of Σ_c from the terms $(y_i - X_i \hat{\boldsymbol{\beta}})(y_i - X_i \hat{\boldsymbol{\beta}})^{\mathrm{T}}$ in which the (u, v)th element is present.

It should be pointed out that when $\boldsymbol{\beta}$ is estimated using (12.45), but with an estimated dispersion matrix, there is no longer any guarantee that the estimator is optimal. If the data can provide a good estimate of the dispersion matrices, then the loss is unlikely to be serious. However, a general dispersion matrix such as Σ_c comprises $\frac{1}{2}k(k+1)$ parameters that need to be estimated, and it would effect a useful saving if the information in the data could be used to estimate fewer parameters. In addition, a general dispersion matrix may well be inappropriate for data collected over time; for example, it may be plausible to expect correlations to decrease as the time between when they were recorded increases. Therefore it may be useful to attempt to provide a model for the dispersion matrix, preferably one using substantially fewer than $\frac{1}{2}k(k+1)$ parameters.

There are various approaches to this task. One is to introduce random effects which induce a particular form for the dispersion matrix. This is the approach outlined in the previous section in the more general setting of multilevel models. An important reference to the application of this approach in the analysis of longitudinal data is Laird and Ware (1982); further details can be found in Chapter 6 of Crowder and Hand (1990).

In some applications the random effects method will be sufficient. However, if the model, for example, for serial measurements of blood pressure on patient i, is as in (12.44), then it may be inadequate to assume that the terms $\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ik}$, are independent and some further modelling may be required. It may be that ε_{ij} could be decomposed as $\varepsilon_{ij} = \zeta_{ij} + \eta_{ij}$. The second of these terms may well be considered to be independent from one measurement occasion to the next, with

constant variance σ_M^2, and would represent items such as measurement error. The first term would have a more complicated dispersion matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ defined in terms of a vector of parameters $\boldsymbol{\theta}$, preferably of low dimension. The dependence between the different elements $\zeta_{i1}, \zeta_{i2}, \ldots, \zeta_{ik}$ measures the genuine serial correlation between the measurements of blood pressure within patient i.

Many different models have been suggested in the literature as candidates for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Only one type is discussed here, the widely used *first-order autoregression model* (see §12.7), which for observations taken at equally spaced intervals, say, times $1, 2, \ldots, k$, has

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = \sigma_W^2 \theta^{|i-j|} \text{ for a scalar } -1 < \theta < 1. \tag{12.46}$$

This form is used because it arises from the following equation for generating the $\zeta_{i1}, \zeta_{i2}, \ldots, \zeta_{ik}$, namely,

$$\zeta_{ij} = \theta \zeta_{i,j-1} + \omega_{ij} \; j = 2, \ldots, k, \tag{12.47}$$

in which each term is related to the previous term through the first part of the equation, but with a random perturbation from the innovation term ω_{ij}. This term comprises independent terms with zero mean and variance $\sigma_W^2(1 - \theta^2)$. It should be noted that the above equation does not specify ζ_{i1} and in order to complete matters it is necessary to supplement the equation with $\zeta_{i1} = \omega_{i1}$ and, if the above equation for $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij}$ is to be obtained, then it is further required to set the variance of ω_{i1} to σ_W^2. This rather clumsy manoeuvre, which appears more natural if the time index in (12.47) is extended indefinitely in the negative direction, is required to obtain a *stationary* autoregression, in which the variance of the ζ term does not change over time, and the correlations depend only on the interval between the occasions concerned. If ω_{i1} had the same variance as the other innovation terms, then a *non-stationary first-order* autoregression would result.

If the matrix in (12.46) is inverted, then the result is a matrix with non-zero entries only on the leading diagonal and the first subdiagonal. This is also true for the dispersion matrix that arises from the non-stationary first-order autoregression. This reflects a feature of the dependence between the observations known as the *conditional independence structure*, which also arises in more advanced techniques, such as graphical modelling (Whittaker, 1990; Cox & Wermuth, 1996). The structure of the inverse dispersion matrix bears the following interpretation. Suppose the blood pressure of a patient had been measured on each day of the week; then, provided the value on Thursday was known, the value on Friday is independent of the days before Thursday. In other words, the information about the history of the process is encapsulated in the one preceding measurement. This reflects the fact that (12.47) is a first-order process, which is also reflected in there being only one non-zero diagonal in the inverse dispersion matrix.

This can be extended to allow second- and higher-order processes. A process in which, given the results of the two previous days, the current observation is then independent of all earlier values is a second-order process, and an rth-order process if the value of the r days preceding the present need to be known to ensure independence. The inverse dispersion matrix would have, respectively, two or r subdiagonals with non-zero entries. Generally, this is referred to as an ante-dependence process, and the two first-order autoregressions described above are special cases. If the total number of observations on an individual is k, then the $(k-1)$th-order process is a general dispersion matrix and the zero-order process corresponds to complete independence. The theory of ante-dependence structures was propounded by Gabriel (1962). An important contribution was made by Kenward (1987), who realized that these complicated models could be fitted by using analysis of covariance, with the analysis at any time using some of the previous observations as covariates.

Although a very useful contribution to the modelling of the covariance of equally spaced data, the ante-dependence models are less useful when the intervals between observations vary. Diggle (1988) proposes a modified form of (12.46) suitable for more general intervals.

The way a covariance model is chosen is also important but is beyond the scope of this chapter. Excellent descriptions of ways to approach empirical modelling of the covariance structure, involving simple random effects and measurement error terms, as well as the serial dependence term, can be found in Diggle *et al.* (1994), especially Chapters 3 and 5.

Generalized estimating equations

Although modelling the covariance structure has considerable logical appeal as a thorough approach to the analysis of longitudinal data, it has some drawbacks. Identifying an appropriate model for the dispersion matrix is often difficult and, especially when there are few observations on each patient or experimental unit, the amount of information in the data on the parameters of the dispersion matrix can be limited. A consequence is that analyses based on (12.44) with estimated dispersion matrices can be much less efficient than might be imagined because of the uncertainty in the estimated dispersion matrices. Another difficulty is that most of the suitable and tractable models are based on the multivariate normal distribution.

An alternative approach is to base estimation on a postulated dispersion matrix, rather than to attempt to identify the correct matrix. The approach, which was proposed by Liang and Zeger (1986) and Zeger and Liang (1986), uses *generalized estimating equations* (GEEs) (see, for example, Godambe, 1991) and has been widely used for longitudinal data when the outcome is categorical. However, it is useful for continuous data and will be described in this context.