The Statistician (1995)
44, *No.* 1, *pp.* 113–135

The analysis of repeated measures: a practical review with examples

By B. S. EVERITT†

Institute of Psychiatry, London, UK

[Received December 1993. Revised June 1994]

SUMMARY

Repeated measures data, in which the same response variable is recorded on each observational unit on several different occasions, occur frequently in many different disciplines. Many methods of analysis have been suggested including *t*-tests at each separate time point and multivariate analysis of variance. In this paper the application of a number of methods is discussed and illustrated on a variety of data sets. The approach involving the calculation of a small number of relevant summary statistics is considered to have advantages in many circumstances.

Keywords: Compound symmetry; Missing values; Multivariate analysis of variance; Repeated measures; Summary measures

1. Introduction

Repeated measures data arise when time sequences of observations of the same dependent variable are made on each of a number of experimental units (usually subjects or patients—in this paper they will be referred to as 'subjects') possibly allocated to one of several treatments. The investigator may also vary systematically the conditions under which the repeated measurements are made, thus introducing one or more within-subject factors into the design. Often the repeated measures include several taken before any treatment starts. The dependent variable may be quantitative or categorical. The following is a small selection of specific examples that have been reported in the increasing literature of repeated measures designs.

(a) A randomized trial of 152 patients with coronary heart disease compared an active drug with a placebo during a 12-month follow-up period. The liver enzyme CPK in serum was measured to study a possible adverse drug effect on the liver. Each patient had three pretreatment measurements taken 2 months before, 1 month before and at randomization, and eight post-treatment measurements taken every 1.5 months after randomization (Frison and Pocock, 1992).

(b) During the grazing season, from spring to autumn, cattle can ingest round-worm larvae, which have developed from eggs previously deposited in the pasture in the faeces of infected cattle. Once infected, an animal is deprived of nutrients and its resistance to other diseases is lowered, which in turn can greatly affect its growth. The monitoring of the effects of a treatment for the disease requires observations to be made throughout the grazing season. In an experiment to compare two methods for controlling the disease, 60 animals were randomly assigned to two treatment groups each of size 30. The animals were put out to pasture at the start of the grazing season, the members of each group receiving one of the two treatments. The weight of each animal was then

†*Address for correspondence*: Department of Biostatistics and Computing, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London, SE5 8AF, UK.

recorded 11 times. The first 10 measurements were made at 2-week intervals and the final measurement was made after a 1-week interval (Kenward, 1987).

(c) Visual acuity was the subject of an investigation in which response times of the eyes to a stimulus were measured. The variable recorded was the time lag between the stimulus (a light flash) and the electrical response at the back of the cortex. Recordings were made for left and right eyes through lenses of four different powers (Crowder and Hand, 1990).

(d) In a comparative clinical trial to investigate which of two drugs made life more tolerable for patients suffering from brain tumours, participants were asked, 'do you still enjoy doing the things you used to?', and their responses in terms of the three categories 'yes', 'sometimes' and 'no' recorded on nine occasions. Here the response variable is categorical (Crowder and Hand, 1990).

The methods used to analyse repeated measures data range from the simple to the complex, with a particular approach often being specific to a particular discipline. In medicine, for example, such data are still largely analysed by a series of t-tests at different time points, whereas in psychology multivariate analysis of variance is frequently employed. The purpose of this paper is to review some of the methods used in practice and to illustrate their application to particular data sets. The theory behind each method will be dealt with only very briefly or omitted entirely. Good sources of theoretical detail are Crowder and Hand (1990), Laird *et al.* (1992), Lindsey (1993) and Diggle *et al.* (1994).

Section 2 considers relatively simple methods including the analysis of separate time points and the use of summary measures. Section 3 describes the use of analysis-of-variance models. Section 4 looks at more complex models, particularly useful when the repeated measures are not taken at the same time for each subject, or when there are missing values, a frequently occurring problem with longitudinal data.

Only the analysis of *quantitative* repeated measures data is discussed in this paper. Categorical variables will be considered in a later paper.

2. Simple methods for the analysis of repeated measures data

A useful initial step in the analysis of repeated measures data is to graph the data in some way. A method often employed, particularly in medical publications, is to plot means by treatment group for every time point. An example of such a plot for the data from the trial of two treatments for the control of intestinal parasites in cattle given in Kenward (1987) is shown in Fig. 1. Standard error bars are frequently attached to such plots, although the resulting diagram can, at times, become quite cluttered. To supplement this 'means' plot, it is usually helpful to produce separate graphs of the responses against time for each subject, differentiating between treatments in some way. Fig. 2, for example, shows the individual growth profiles for the 30 animals on each treatment. When the number of individual curves is large it may be more useful to plot a small number of representative curves. A suggestion due to Jones and Rice (1992) may be helpful here. The idea is to use principal components analysis to select a small number of the growth curves. Specifically, those curves corresponding to the units with the minimum, maximum and median principal component scores on the first one or two principal components will generally give an adequate summary of all the curves. For the cattle growth data the first principal component is simply a weighted average of the weights on each of the 11 occasions, and the second component is essentially a contrast of the weights on the first eight occasions, with those on the last three. The summary curves chosen according to the individual principal component scores are shown in Figs 3 and 4. These curves highlight the apparent decline in weight on the last two or three occasions for those animals given treatment B.

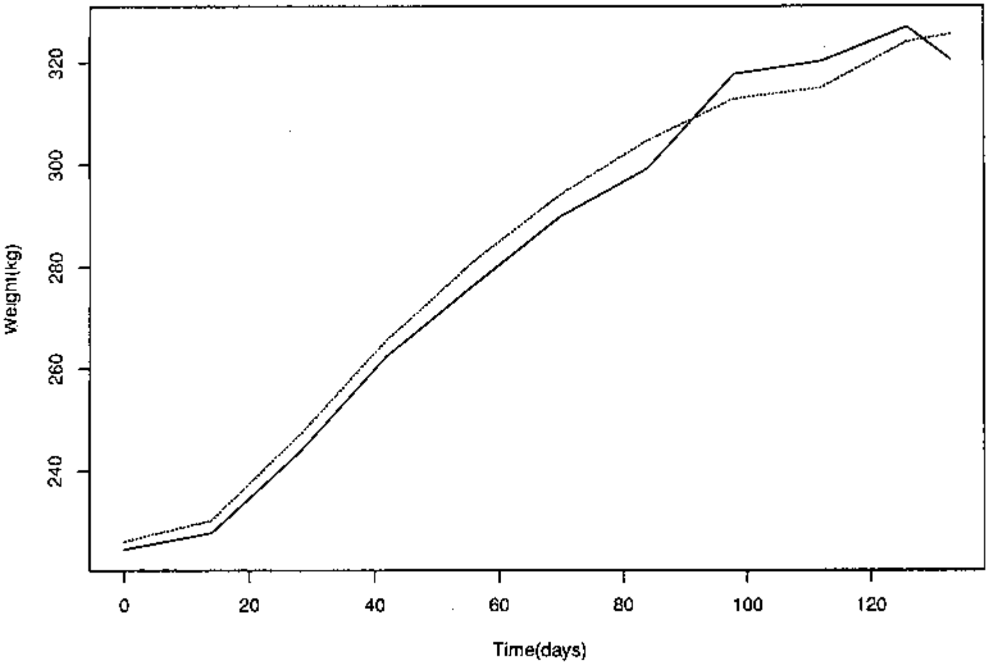

Fig. 1. Plot of treatment means for the cattle growth data: ······, treatment A; ——, treatment B

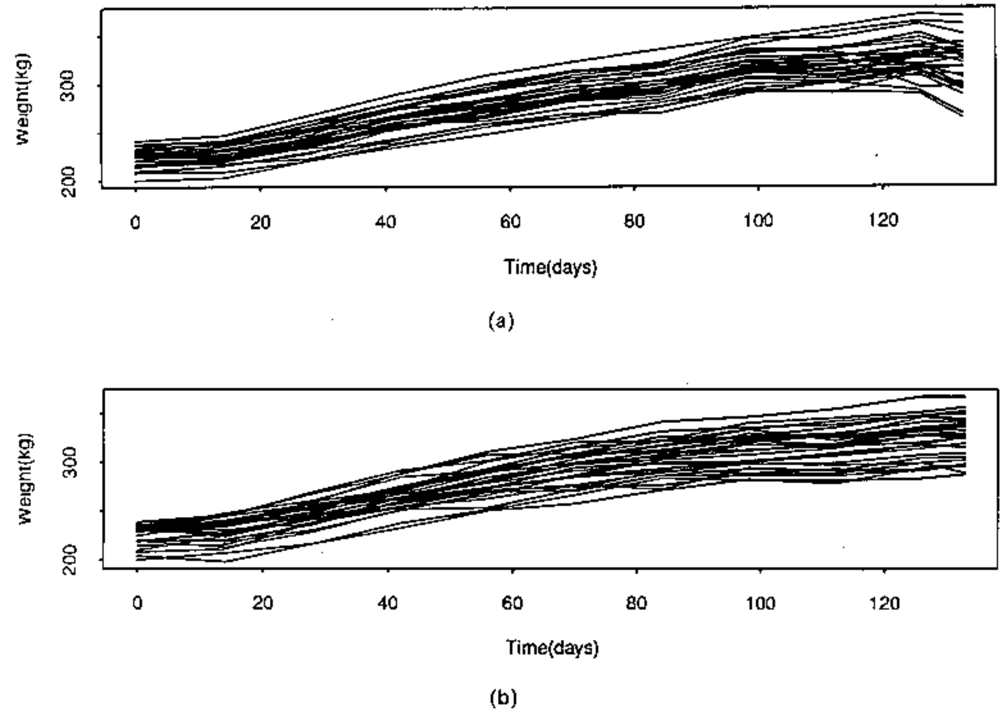

Fig. 2. Individual growth profiles in each treatment group for the cattle growth data: (a) treatment B; (b) treatment A

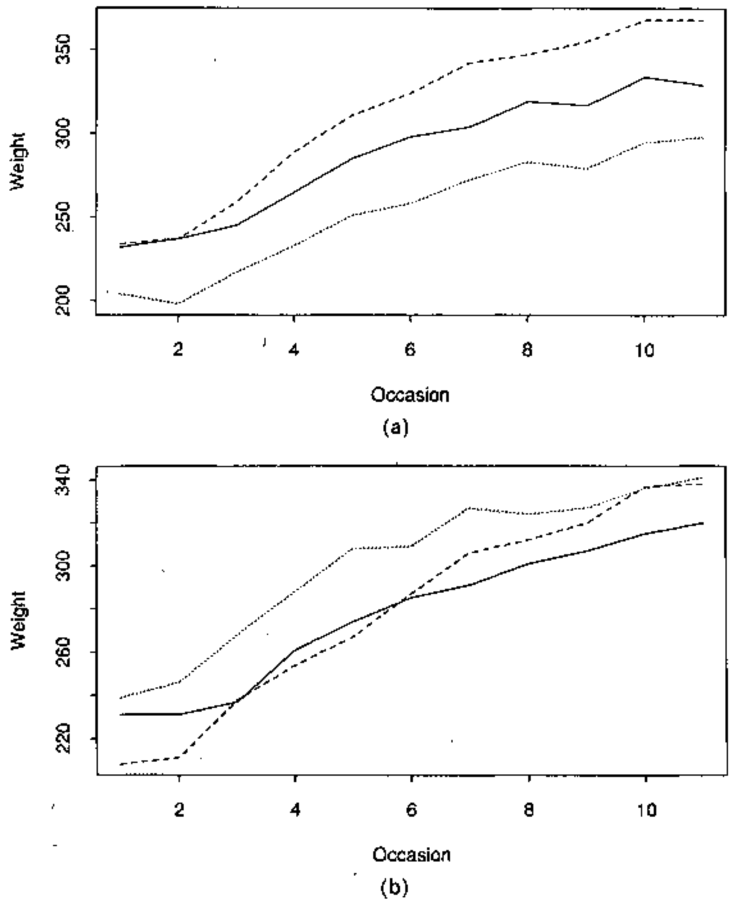

Fig. 3. Summary curves chosen by the use of principal component scores (treatment A; ———, median; · · · · · ·, minimum; - - - - -, maximum): (a) principal component 1 curves; (b) principal component 2 curves

A commonly used method of analysis for repeated measures that involve a number of treatment groups, particularly in medical and related research, is to compare the groups at each time point, by using either *t*-tests or some nonparametric equivalent. Table 1, for example, shows the results of using the procedure on the cattle weight data. None of the *t*-tests show any evidence of a treatment difference.

Finney (1990) suggested that this approach may be quite useful if the occasions are few and the intervals between them are large. In general, however, there are convincing arguments against such multiple tests. The first is that the tests are clearly *not* independent and so their interpretation is difficult. Simply assuming that the tests give independent information about group differences is clearly not sensible, as is demonstrated by considering what would happen if the repeated measurements were made more frequently. The number of significance tests performed would rise accordingly, but the increased information about the difference between the treatments is likely to be very small. Repeated testing also assumes that each time point is of separate interest in its own right. This is unlikely in most cases; the real interest is likely to be in something more global. The separate significance tests do not give an overall answer to whether or not there is a treatment difference and provide no single estimate of the treatment effect.

A more relevant, but still relatively straightforward, approach to the analysis of repeated measures data is that involving the use of *summary measures*, and sometimes known as

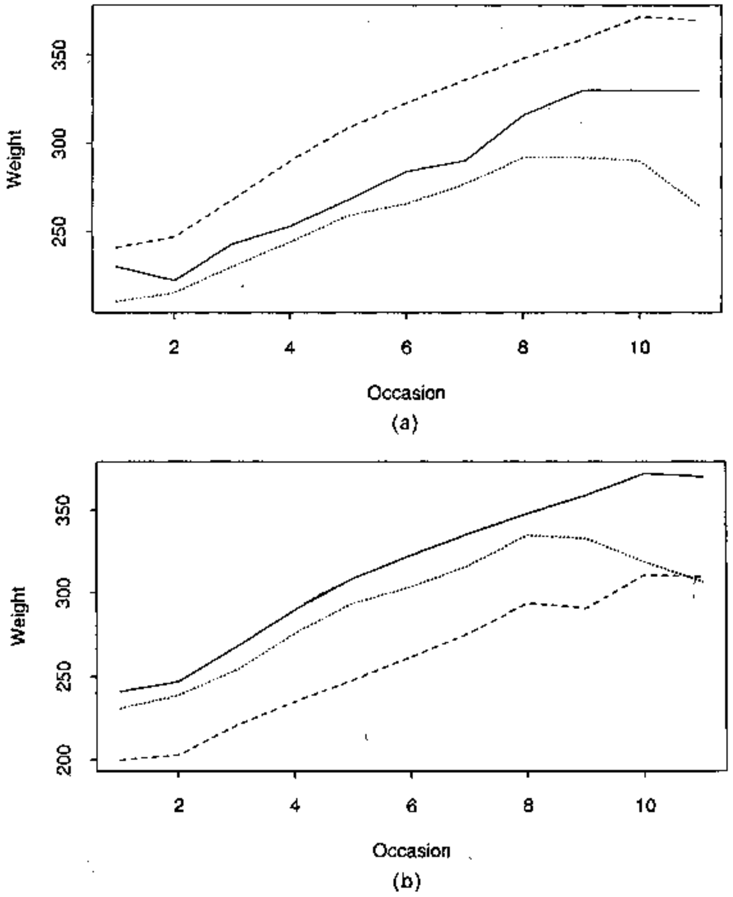

Fig. 4. Summary curves chosen by the use of principal component scores (treatment B; ———, median; · · · · · ·, minimum; - - - -, maximum): (a) principal component 1 curves; (b) principal component 2 curves

response feature analysis. Here the responses for each subject are used to construct a single number that summarizes some aspect of the subject's response profile. (In some cases more than a single summary measure may be used.) The summary measure to be used needs to be chosen before the analysis of the data and should, of course, be relevant to the particular questions of interest in the study. Matthews *et al.* (1990) gave a list of potentially useful summary measures, reproduced here in Table 2. In clinical trial work, Frison and Pocock (1992) argued that the average response to treatment over time is often likely to be the most relevant summary of the repeated measurements.

Having identified a suitable summary measure, the analysis of treatment differences generally involves the application of a simple univariate test (usually a t-test or nonparametric equivalent) to the single measure now available for each subject. Table 3 shows the results of the analysis of several different summary measures calculated from Kenward's cattle weight data. Once again there is no evidence of any treatment effect. (Diggle *et al.* (1994) consider other summary measures that do show a treatment difference.)

Pretreatment measures, if available, can be used in association with the response feature approach in several ways. If, for example, the average response to treatment over time is the chosen summary, Frison and Pocock (1992) suggest three possible methods of analysis:

TABLE 1
t-tests for each time point for cattle growth data†

Time (days)		*A*	*B*	*t*	*P*
0	Mean	226.20	224.63		
	SD	10.27	10.20	−0.59	0.56
14	Mean	230.33	227.90		
	SD	12.45	10.41	−0.82	0.41
28	Mean	246.87	243.53		
	SD	12.85	12.13	−1.03	0.30
42	Mean	265.63	262.47		
	SD	13.60	14.15	−0.88	0.38
56	Mean	281.13	276.43		
	SD	15.56	14.75	−1.20	0.23
70	Mean	294.33	288.07		
	SD	17.35	19.45	−1.32	0.19
84	Mean	304.73	299.23		
	SD	17.51	15.75	−1.28	0.21
98	Mean	312.87	317.67		
	SD	18.46	15.30	1.10	0.28
112	Mean	315.03	320.20		
	SD	19.87	17.01	1.08	0.29
126	Mean	324.07	326.93		
	SD	21.68	20.12	0.53	0.60
133	Mean	325.47	320.50		
	SD	21.08	24.39	−0.84	0.40

†SD, standard deviation.

(a) *post-treatment means* (POST)—a simple analysis using the mean for each subject's post-treatment responses as the summary measure;
(b) *mean changes* (CHANGE)—a simple analysis of each subject's difference between the mean of post-treatment responses and the mean of base-line measurements, the latter often consisting of a single base-line value per subject;
(c) *analysis of covariance* (ANCOVA)—between-subject variations in base-line measurements are taken into account by using the mean of the base-line values for each subject, as a covariate in a linear model for the comparison of post-treatment means.

TABLE 2
Some summary measures as given in Matthews *et al.* (1990)

Type of data	*Question of interest*	*Summary measure*
Peaked	Is overall value of outcome variable the same in different groups?	Overall mean (equal time intervals); area under curve (unequal time intervals)
Peaked	Is maximum (minimum) response different between groups?	Maximum (minimum) value
Peaked	Is time to maximum (minimum) response different between groups?	Time to maximum (minimum) response
Growth	Is rate of change of outcome variable different between groups?	Regression coefficient
Growth	Is eventual value of outcome variable the same between groups?	Final value of outcome measure or difference between last and first values, or percentage change between first and last value
Growth	Is response in one group delayed relative to the other?	Time to reach a particular value (e.g. a fixed percentage of base-line)

TABLE 3
Analysis of summary measures for cattle growth data†

		A	*B*	*t*	*P*
Mean	Mean	284.24	282.51		
	SD	14.94	14.02	−0.45	0.64
Maximum	Mean	326.43	330.23		
	SD	20.83	18.27	0.75	0.46
Linear slope	Mean	0.80	0.83		
	SD	0.15	0.13	0.89	0.38

†SD, standard deviation.

The results of applying each of these methods to the cattle weight data, regarding the observation at time 0 as pretreatment, are shown in Table 4. All three approaches indicate that the data show no evidence of a treatment effect. (In any detailed analyses of these data, assumptions such as normality and equality of regression slopes within groups would, of course, need to be checked.)

Frison and Pocock (1992) compared the three approaches to the analysis of repeated measures data when pretreatment values are available. With a single pretreatment recording, they found that analysis of covariance is more powerful than both analysis of change scores and analysis of post-treatment means only, except when the correlations between the repeated measures are small. Using the mean of several pretreatment measures as a covariate makes the analysis of covariance even more efficient if there are substantial correlations between the repeated measures. The differences between the three approaches can be illustrated concisely by comparing power curves calculated by using the formulae given in Frison and Pocock (1992). Figs 5–7 show some examples for the situation with two treatment groups, two pretreatment and four post-treatment observations. (The correlations between pairs of repeated measures are assumed *equal* in the calculation of these power curves.) The sample size needed to achieve a particular power to detect a standardized difference of 0.5 is always lower with analysis of covariance, and in some cases *substantially* lower. As the correlation increases, CHANGE becomes less inferior to method ANCOVA. Both become substantially better than POST. With a correlation of 0.2, CHANGE is seen to be less efficient than simply dealing with post-treatment values. (When the correlation is 0 the ANCOVA and POST methods are almost equivalent.)

TABLE 4
Analysis of cattle growth data†

	A	*B*	*t*	*P*
(*a*) *POST*				
Mean	286.11	284.71		
SD	15.39	14.16	0.35	0.72
(*b*) *CHANGE*				
Mean	59.91	60.08		
SD	11.91	10.12	0.00	0.95
(*c*) *ANCOVA*				
Adjusted means	285.36	285.47	0.00	0.97

†SD, standard deviation.

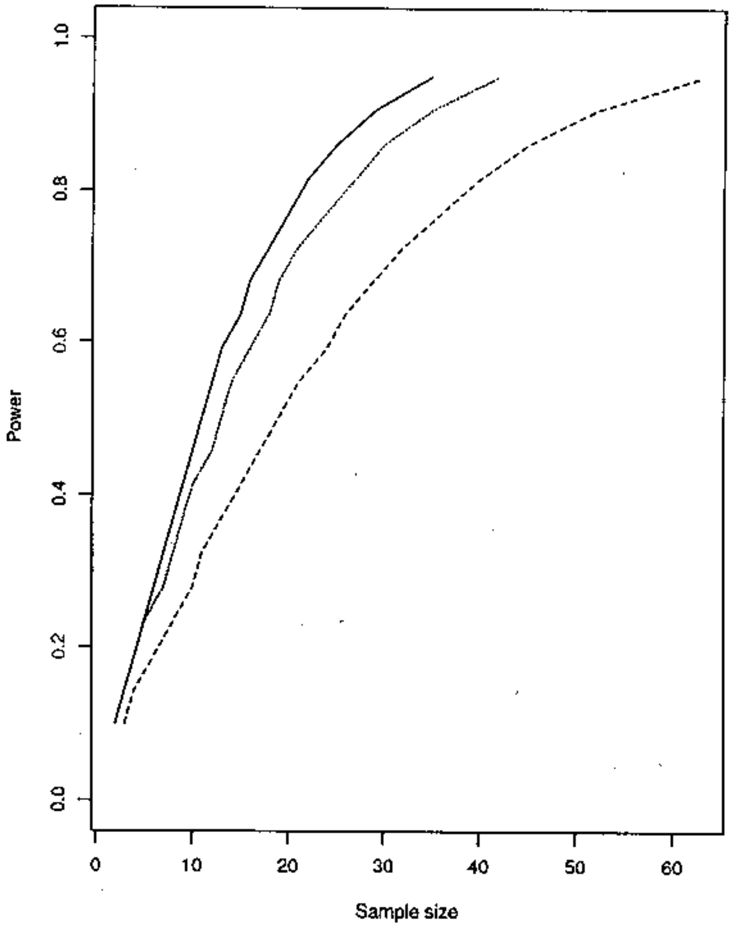

Fig. 5. Power curves for three methods of analysing mean response values ($\rho = 0.2$): ———, ANCOVA; · · · · · ·, POST; - - - - -, CHANGE

The use of summary measures in the analysis of repeated measures data has several advantages. Three listed by Matthews (1993) are

(a) an appropriate choice of summary measures ensures that the analysis is focused on relevant and interpretable aspects of the data.
(b) the method is statistically valid and
(c) to an extent, missing and irregular observations can be accommodated (this point will be the subject of further consideration in Section 4).

There are, however, some possible disadvantages with such an approach. First it may be difficult to specify in advance an appropriate and relevant summary measure. Matthews *et al.* (1990), however, pointed out that this may, in fact, be an *advantage* since it might encourage researchers to think about the features of the data that will be of most interest to them when designing the study, rather than simply posing the rather vague question, 'how do the groups differ?'. A further statistical problem in using the response feature approach to the analysis of repeated measures data arises from assuming that the individual summary measures in each group are identically distributed. In Table 3, for example, one of the analyses involves a t-test for the difference in the treatment group means of the linear regression slopes

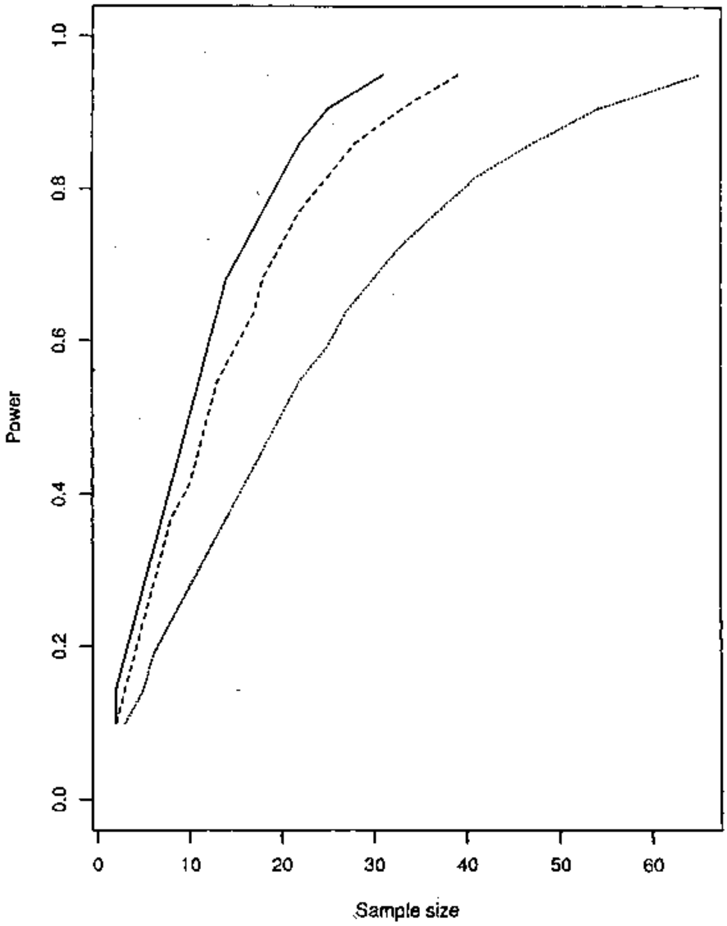

Fig. 6. Power curves for three methods of analysing mean response values ($\rho = 0.5$): ———, ANCOVA; · · · · · ·, POST; - - - - -, CHANGE

of each calf. This t-test assumes that the slopes have identical normal distributions within groups. But, if there is natural variation in the slopes between individuals even in the same group, then such an assumption is clearly not valid. The distributions of the individual slopes will have different means.

Ignoring that a summary measure is *not* a single observation but an aggregate of data from a subject may, on occasions, involve a loss of information that could be used to allow a more efficient analysis. Matthews (1993) addressed this problem and described some *ad hoc* procedures that might be useful in particular circumstances. Gornbein *et al.* (1992) also considered how the summary measure approach can be made more efficient and considered several models in which some form of weighting is introduced. This will be of particular importance if the configuration of repeated measures for each subject varies, resulting in the individual summary measures being estimated with differing precisions.

3. Analysis-of-variance methods for repeated measures data

In many areas, particularly psychology, repeated measures data often arise from designed experiments. The traditional analysis of such data in these disciplines is by univariate analysis

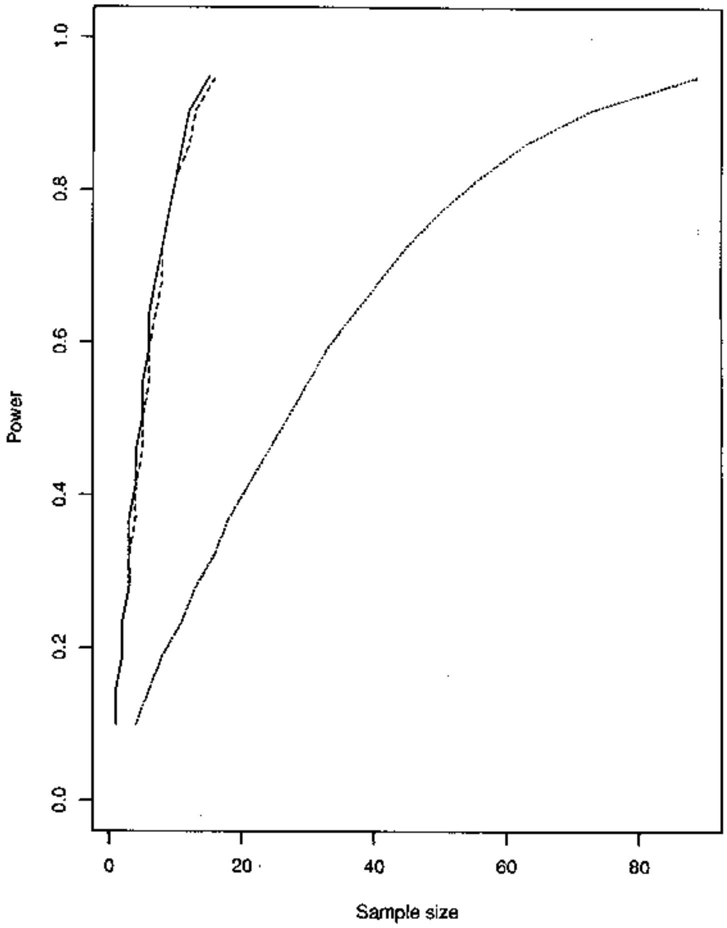

Fig. 7. Power curves for three methods of analysing mean response values ($\rho = 0.8$): ———, ANCOVA; · · · · · ·, POST; - - - - -, CHANGE

of variance in which the within-subject factors are considered part of the experimental design as reflected in the summary table. The usual models behind such analyses are described in detail in Winer (1971) but in general involve the sum of *fixed* and *random* effects. The former correspond to design variables and their interactions. The latter arise from regarding subjects as a sample from some population. Subject effects are taken to be random as are the interactions of the subject and design variables. For Kenward's cattle weight data this approach leads to the analysis of variance shown in Table 5. The terms corresponding to the main effect of time and the group × time interaction are both highly significant. The first of these simply indicates the growth of the animals over the time period considered and is not of great interest. The significant interaction, however, suggests that the growth profiles in the two groups are not parallel and that, consequently, the two treatments involved lead to different patterns of growth. Fig. 1 indicates that this results largely from the behaviour of the curves after day 84. Further interpretation of this interaction would need more substantive knowledge of investigation but clearly here the analysis-of-variance approach leads to a different conclusion from that when the summary measure procedure is used, at least for the summary measures considered.

TABLE 5
Analysis of variance for cattle growth data

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*	*F*	*P*
Group	497.47	1	497.47	0.22	0.64
Error	133920.55	58	2308.97		
Time	845724.68	10	84572.47	1191.53	<0.0001
Group × time	2558.71	10	255.87	3.60	0.0001
Error	41167.15	580	70.98		

To illustrate the analysis-of-variance approach to repeated measures data further, in an area where it is more likely to be employed, the data shown in Table 6 taken from Broota (1989) will be used. These data arise from an experiment in which two groups of subjects are required to read two types of word under three cue conditions. Details are given in Pahwa and Broota (1981). Here the repeated measures arise from the crossing of two experimental factors. The resulting analysis of variance is shown in Table 7. The main effects of type and cue are both highly significant and the type × cue interaction appears to be of possible interest, a point which will be examined later.

TABLE 6
Observations obtained from a 2 × 2 × 3 factorial experiment with repeated measures on the last two factors†

Subject	b_1 *(form)*			b_2 *(colour)*		
	$c_1(N)$	$c_2(C)$	$c_3(I)$	$c_1(N)$	$c_2(C)$	$c_3(I)$
a_1 *(FI)*						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	197	158	159	169
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
a_2 *(FD)*						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	180	187	139	151	163

†FI, field independent; FD, field dependent; N, normal; C, congruent; I, incongruent.

TABLE 7
Analysis of variance for the data in Table 6

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*	*F*	*P*
Groups	18906.25	1	18906.25	2.56	0.12
Error	162420.08	22	7382.73		
Type	25760.25	1	25760.25	12.99	0.0016
Groups × type	3061.78	1	3061.78	1.54	0.23
Error	43622.30	22	1982.83		
Cue	5697.04	2	2848.52	22.60	<0.0001
Groups × cue	292.62	2	146.31	1.16	0.32
Error	5545.00	44	126.02		
Type × cue	345.37	2	172.69	2.66	0.08
Groups × type × cue	90.51	2	45.26	0.70	0.50
Error	2860.78	44	65.02		

The F-tests in Tables 5 and 7 are valid only if a particular set of conditions holds. Normality of the response variable is, of course, one necessary condition. Of more critical importance are the conditions relating to the variances and covariances of the repeated measures. Explicitly, the covariance matrix of the repeated measures must be such that the elements on the main diagonal, the variances of the repeated measures, are equal, and the off-diagonal elements, the covariances of each pair of repeated measures, are also equal. A covariance matrix of such form is said to have *compound symmetry*. Additionally the validity of the F-tests in Tables 5 and 7 requires that this covariance matrix is the same in each treatment group.

Compound symmetry is a special case of a more general situation under which the simple F-tests are valid. This more general condition is known as *sphericity* or *circularity* and relates to the covariance matrix of a set of $p-1$ orthonormal contrasts between the p repeated measures. Specifically this covariance matrix must be a scalar multiple of the identity matrix for the F-tests to be valid. A detailed account of sphericity and compound symmetry is given in Crowder and Hand (1990).

If the sphericity condition does not hold then the univariate F-tests of Tables 5 and 7 are not correct, and their use will lead to an increase in the size of the type 1 error. Various tests of the condition are available (see, for example, Mauchly (1940)), but they are of limited practical use because of their known sensitivity to non-normality, to which the F-tests in question are relatively robust. In many situations, however, the compound symmetry assumption is likely to be questionable *a priori*. It is very probable that, with repeated measures data, observations closer together in time will be more highly correlated than those separated by a greater time interval. Table 8, for example, shows the correlation matrices for each treatment group in the cattle growth data. Neither matrix appears to have the compound symmetry form, the correlations of measurements close in time being noticeably higher than those with wider separation. The Mauchly sphericity test of the pooled within-group matrix gives a P-value less than 0.0001. Clearly the F-values for these data, given in Table 5, are not strictly valid.

If sphericity does not hold, the univariate F-tests may be adapted to non-sphericity by estimating a correction factor that measures the departure from sphericity. This correction factor, which is a function of the variances and covariances of the repeated measures, is defined explicitly in Crowder and Hand (1990); it is used to reduce the degrees of freedom of the F-tests associated with the within-subjects part of the analysis-of-variance table. Two methods of estimating the correction factor have been suggested, one by Greenhouse and Geisser (1959)

TABLE 8
Correlation matrices for cattle growth data

	Occasion										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
(a) Treatment A											
1	1.000										
2	0.825	1.000									
3	0.764	0.907	1.000								
4	0.659	0.844	0.925	1.000							
5	0.635	0.803	0.879	0.941	1.000						
6	0.548	0.736	0.826	0.905	0.936	1.000					
7	0.524	0.628	0.748	0.825	0.872	0.935	1.000				
8	0.530	0.667	0.771	0.837	0.893	0.959	0.932	1.000			
9	0.512	0.599	0.707	0.767	0.836	0.919	0.933	0.969	1.000		
10	0.476	0.584	0.699	0.734	0.799	0.885	0.883	0.943	0.964	1.000	
11	0.479	0.551	0.679	0.713	0.773	0.849	0.864	0.924	0.958	0.984	1.000
(b) Treatment B											
1	1.000										
2	0.862	1.000									
3	0.835	0.944	1.000								
4	0.685	0.893	0.930	1.000							
5	0.673	0.842	0.881	0.945	1.000						
6	0.540	0.681	0.763	0.781	0.831	1.000					
7	0.613	0.781	0.822	0.913	0.935	0.780	1.000				
8	0.637	0.809	0.838	0.922	0.912	0.783	0.953	1.000			
9	0.644	0.773	0.790	0.869	0.905	0.769	0.907	0.951	1.000		
10	0.481	0.651	0.668	0.776	0.778	0.632	0.759	0.782	0.834	1.000	
11	0.443	0.569	0.622	0.723	0.684	0.563	0.715	0.713	0.757	0.924	1.000

and one by Huynh and Feldt (1976). For many data sets the estimates are likely to be very similar. When sphericity holds, the correction factor takes the value 1, and its smallest possible value is $1/(p-1)$. Greenhouse and Geisser (1959) suggested this lower limit in all cases, thus avoiding the need to estimate the correction factor at all. Such a strategy is, however, very conservative.

The correction factor approach can be illustrated on both the cattle growth data from Kenward (1987) and the reaction time data of Broota (1989). For the former the estimated correction factor using the method described in Greenhouse and Geisser (1959) is 0.2767, and using the method suggested by Huynh and Feldt (1976) it is 0.2970. Using the former, the degrees of freedom for the F-test of differences between time points become $10 \times 0.2767 = 2.8$ and $580 \times 0.2767 = 160.5$. The associated P-value remains very small. For the group × time interaction, the degrees of freedom of the relevant F-test also become 2.8 and 160.5. The P-value increases from 0.0001 to 0.0173. If the conservative strategy of using the value $1/(p-1) = 0.1$ as the correction factor had been employed, the P-value for the interaction would be above 0.05, and the interaction effect deemed non-significant.

The sphericity test for the Broota data in Table 6 has a P-value 0.03 and the estimated correction factor for the tests involving the 'cue' factor is 0.96 (Huynh–Feldt), and, for the tests involving the factor 'type', takes the value 0.86. Here departure from the sphericity assumption is less pronounced than for the cattle growth data, and adjustment of the F-tests given in Table 7 by the relevant correction factor makes little difference to the associated P-values.

An alternative to the use of correction factors when the sphericity assumption does not hold is to adopt a *multivariate* approach to the repeated measures. The main advantage of this method is that no assumptions are made about the form of the covariance matrix of the repeated measures, although this covariance matrix is still required to be the same in each treatment group. The disadvantage of the multivariate approach to the analysis of repeated measures is often stated to be its relatively low power when the sphericity assumption is, in fact, valid (Crowder and Hand, 1990; Rouanet and Lepine, 1970). Davidson (1972), however, compared the power of the two tests when compound symmetry holds and concluded that the multivariate test is nearly as powerful as the univariate test when the number of observations exceeds the number of repeated measures by 20 or more.

The multivariate procedure involves testing on a set of transformed variables representing the within-subject differences of each within-subject factor and their interactions. (If a factor has only two levels, the univariate and multivariate approaches are equivalent.) Each of the transformed variables has some within-cell variance that can be used as the error term in the analysis of that variable. In some cases each of the transformed variables is examined individually to give what Hand and Taylor (1987) call a *multiple univariate analysis*. In others the transformed variables are assessed simultaneously by using a multivariate test. The hypothesis that the means of a set of transformed variables representing a within-subject factor, or an interaction between within-subject factors, are zero can be tested by using Hotelling's T^2-statistic. The interactions of within-subject factors with treatment groups are tested by using the same statistic if the number of groups is 2, or with one or other of a variety of possible test statistics if there are more than two groups.

For the cattle growth data an obvious set of transformed variables is the *orthogonal polynomials*, representing linear, quadratic, cubic, etc., effects over time. A multiple univariate analysis of these variables is shown in Table 9. Each effect is tested by using its own particular error term. In this analysis no multivariate test criteria have been used. For these data it is unlikely that the higher order polynomial terms would be of any interest, and in practice several would be combined.

The results of applying the multivariate approach to the Broota data are shown in Table 10. The most notable difference between this analysis and the univariate analysis given earlier (see Table 7) concerns the test for the type × cue interaction. Using Hotelling's T^2, this interaction is found to be significant beyond the 5% level; the univariate tests, both adjusted and unadjusted, are not significant. Some insight into the reason for this difference can be obtained from the pair of type × cue interaction effects for each individual. These are calculated as follows (using the labelling nomenclature in Table 6):

$$\text{interaction effect one} = b_1c_1 - b_1c_2 - b_2c_1 + b_2c_2;$$

$$\text{interaction effect two} = b_1c_2 - b_1c_3 - b_2c_2 + b_2c_3.$$

A scatterplot of these effects is shown in Fig. 8. Separate t-tests that the mean effects are zero give $t = 1.97$, $P = 0.06$, and $t = -0.10$, $P = 0.92$. Fig. 8 shows that there are two subjects with very large values for the second effect (subjects 5 and 18). When these subjects are removed, Hotelling's $T^2 = 14.27$ with a P-value of 0.005, the two separate t-tests now have values 3.24 ($P = 0.004$) and -1.29 ($P = 0.21$) and the type × cue interaction in the analysis of variance has a P-value of 0.0042. It appears that these two 'outliers' are the reason that the original univariate and multivariate approaches lead to different conclusions.

4. More complex models for repeated measures data

The repeated measures data most appropriately analysed by the methods described in the previous section are those from designed experiments where all subjects have the same number of observations measured at equivalent time intervals. In many cases, however, particularly longitudinal data arising from clinical trials, the number of repeated measures on each subject may not be the same, or they may not be taken at the same interval. Additionally gaps in the data may occur because subjects cease to comply with their assigned treatment and drop out of the study, or they simply fail to attend an appointed visit. An investigator may choose

TABLE 9
Analysis of variance of cattle growth data using orthogonal polynomials†

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*	*F*	*P*
Group	497.47	1	497.47	0.12	0.64
Error	133920.55	58	2308.97		
Time1	816222.84	1	816222.84	2114.13	0.000
Time1 × group	307.54	1	307.54	0.80	0.38
Error	22392.68	58	386.08		
Time2	19092.27	1	19092.27	154.09	0.000
Time2 × group	117.82	1	117.82	0.95	0.33
Error	7186.43	58	123.90		
Time3	6647.44	1	6647.44	115.93	<0.0001
Time3 × group	545.73	1	545.73	9.52	0.0031
Error	3325.81	58	57.34		
Time4	1379.19	1	1379.19	49.76	<0.0001
Time4 × group	1044.19	1	1044.19	37.67	<0.0001
Error	1607.61	58	27.72		
Time5	1161.75	1	1161.75	86.69	<0.0001
Time5 × group	283.25	1	283.25	21.14	<0.0000
Error	777.25	58	13.40		
Time6	118.32	1	118.32	4.74	0.0336
Time6 × group	24.95	1	24.95	1.00	0.3216
Error	1448.02	58	24.97		
Time7	60.95	1	60.95	4.78	0.0328
Time7 × group	38.12	1	38.12	2.99	0.0891
Error	739.36	58	12.75		
Time8	582.14	1	582.14	25.64	<0.0001
Time8 × group	57.77	1	57.77	2.55	0.12
Error	1316.64	58	22.70		
Time9	312.96	1	312.96	29.46	<0.0001
Time9 × group	94.28	1	94.28	8.88	0.004
Error	616.13	58	10.62		
Time10	146.81	1	146.81	4.85	0.03
Time10 × group	45.05	1	45.05	1.49	0.23
Error	1757.22	58	30.30		

†Time1 to time10 represent orthogonal polynomial components, so that time1 is the linear trend, time2 the quadratic trend etc.

to extrapolate, fill in or *impute* missing values by using any one of a variety of methods that have been suggested. Appropriate mean values obtained from the non-missing observations might, for example, be inserted. For several reasons, Gornbein *et al.* (1992) do not recommend this approach. Another popular method of dealing with the missing data problem, particularly in the pharmaceutical industry, is to carry the last recorded value of a subject forwards to produce a 'complete' set of repeated measures. According to Heyting *et al.* (1992), the usefulness of this method is very limited since it makes very unlikely assumptions about the data. Imputation methods need to be carefully chosen to avoid biased estimates from the filled-in data. Also, as Gornbein *et al.* (1992) stressed, imputation *invents* data, and analysing

TABLE 10
Multivariate analysis of the Broota data

Effect	T^2	*F*	*Degrees of freedom*	*P*
Cue	44.67	21.32	2, 21	<0.0001
Cue × group	3.18	1.52	2, 21	0.24
Type × cue	10.41	4.97	2, 21	0.02
Group × type × cue	1.49	0.71	2, 21	0.50

filled-in data as if they were complete leads to overstatement of precision, i.e. standard errors are underestimated, stated *P*-values are too small and confidence intervals do not cover the true parameter at the rate stated.

The preferred approach to irregular or incomplete repeated measures data is to deal with them in the context of a suitable model and to use the method of maximum likelihood to estimate parameters and their standard errors. The advantages of this approach are listed in Gornbein *et al.* (1992) and are as follows:

(a) maximum likelihood estimates are principled in that they have known statistical properties (consistency, large sample efficiency) under the assumed model, which can be clearly specified and subjected to model criticism;
(b) maximum likelihood estimation does not require a rectangular data matrix and hence deals directly with problems of missing data;
(c) estimates are asymptotically efficient under the assumed model;

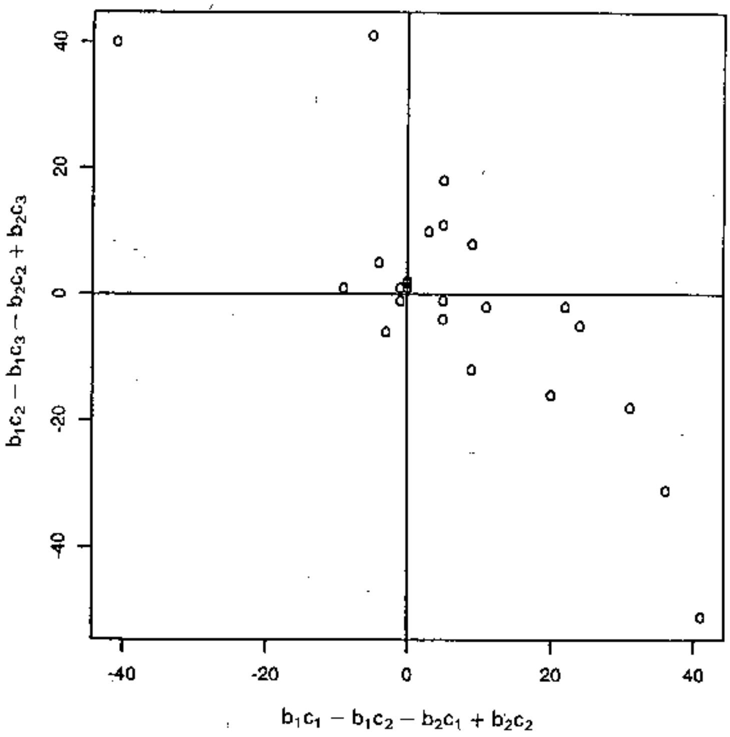

Fig. 8. Scattergram of type × cue effects for the Broota data

(d) standard errors of parameter estimates based on the observed or expected information matrix are available and automatically take into account the fact that the data are incomplete.

The main disadvantages of this approach are

(a) maximum likelihood estimation requires the specification of a full statistical model for the data and results may be vulnerable to departures from model assumptions such as normality and
(b) maximum likelihood inferences are based on large sample theory and hence may be unsuitable for small data sets.

Diggle (1988) lists some desirable features for a general method for the analysis of repeated measures data. These include the following:

(a) the specification of the mean response profile needs to be sufficiently flexible to reflect both time trends within each treatment group and differences in these time trends between treatments;
(b) the specification of the covariance structure within each time sequence should be flexible, but economical;
(c) the method of analysis should accommodate virtually arbitrary patterns of irregularly spaced time sequences within subjects.

Very general regression-type models incorporating these and other features are given in Laird *et al.* (1992) and Gornbein *et al.* (1992). An essential feature of these models is that the parameters of the covariance matrix, which can be allowed to have a variety of forms, are estimated separately from the other parameters in the model. Although the covariance structure is not of direct interest, Diggle (1988) suggests that overparameterization will lead to inefficient estimation and potentially poor assessment of standard errors for estimates of the mean response profiles, whereas too restrictive a specification will invalidate inferences about the mean response profiles when the assumed covariance structure does not hold. In many repeated measures examples the model for the covariances will need to allow for *non-stationarity* with changes in variance across time being particularly common.

To illustrate the application of these more general models, data from a trial of oestrogen patches in the treatment of post-natal depression will be used. Details of the study are given in Kumar *et al.* (1993), but essentially women were randomly allocated to two groups: the members of one group received the active drug, whereas the others received a placebo. The main response variable was a composite measure of depression, this being observed on two occasions before treatment and on six occasions during treatment. The number of women taking part in the trial was 61, of whom 27 were given the placebo and 34 oestrogen.

Of the 61 women in the trial, 45 had no missing values. Table 11 shows the patterns of the incomplete observations. There is no strong evidence that the missing values are related to treatment. A plot of the group means here (Fig. 9) shows a general decline of the depression scores after treatment in both the active and the placebo groups. During the last two occasions the depression scores begin to level off. The two within-treatment-group covariance matrices of the repeated measures, based on women with complete data, are shown in Table 12. It appears that a model for the covariance structure that allows for both different variances on the six visits and different covariances for different pairs of the repeat measures will be necessary here.

The initial model fitted contained a main effect for group and a linear trend for time. The covariance matrix of the six repeated measurements was allowed to be unstructured. The results are shown in Table 13. Both the treatment effect and the linear time effect are highly significant. Next the mean of the two base-line observations was considered for inclusion as a covariate. The estimated regression coefficient was 0.38 with a standard error of 0.16. Consequently the base-line mean was added to the initial model. Finally a quadratic effect

TABLE 11
Pattern of missing (M) values in the treatment trial for post-natal depression

Subject	*Visit*					
	1	*2*	*3*	*4*	*5*	*6*
Placebo group						
1	M			M	M	M
2		M	M	M	M	M
3		M	M			M
4			M	M	M	M
5			M	M	M	M
6	M			M	M	M
7			M	M	M	M
8			M	M	M	M
9		M	M	M	M	M
10		M	M	M	M	M
Active group						
11		M	M			M
12			M	M	M	M
13	M			M	M	M
14		M	M	M	M	M
15	M	M	M			M
16				M	M	M

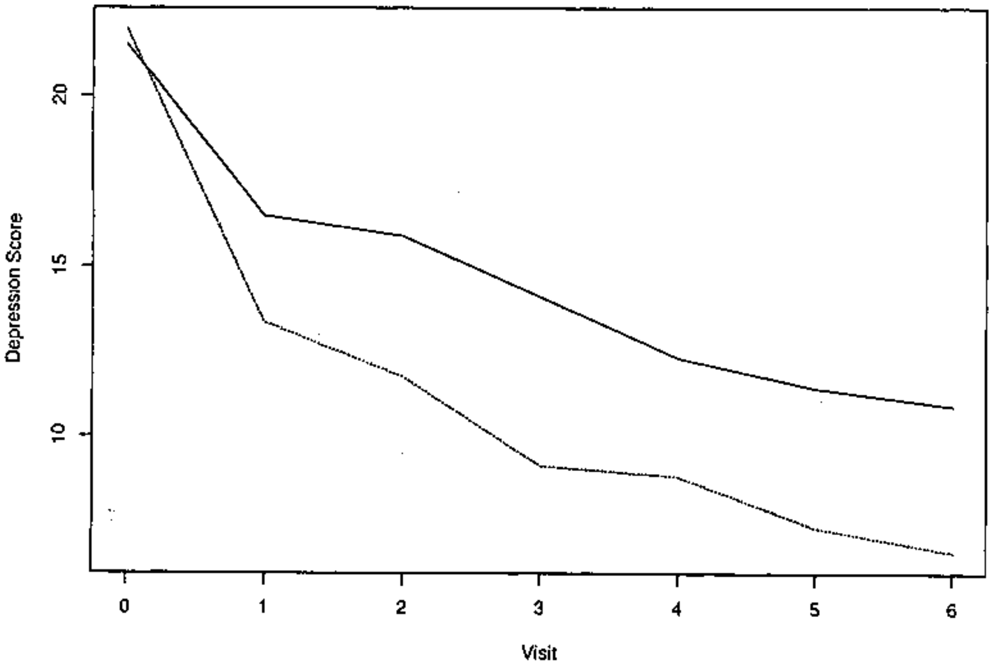

Fig. 9. Plot of group means for the trial of oestrogen in the treatment of post-natal depression (visit 0 is the mean of two pretreatment values): ———, placebo group; ·······, active group

TABLE 12
Covariance matrices for the placebo and active groups in the oestrogen trial (based on complete observations only)

	1	*2*	*3*	*4*	*5*	*6*
Placebo group						
1	32.01					
2	22.06	36.93				
3	15.94	22.14	24.75			
4	13.88	22.84	20.65	34.21		
5	9.51	15.83	19.76	21.74	19.7	
6	−0.18	11.60	15.53	17.87	16.85	21.92
Active group						
1	30.98					
2	11.96	40.11				
3	12.14	32.05	29.77			
4	6.84	22.69	20.44	21.78		
5	8.08	28.66	25.29	23.24	32.96	
6	7.57	23.77	20.61	18.82	25.21	22.37

for time was added to the model and, although not quite significant at the 5% level, was retained. The parameter estimates etc. for this model are shown in Table 14. (Other models which allow for differences in the linear trend in each group etc. were considered but found not to be needed.)

Having decided on a reasonable model for the data by assuming an unstructured covariance matrix, some attention now needs to be given to whether some form of restricted covariance model for the repeated measures might be adequate. Table 15 shows a comparison of log-likelihoods for the fitted model with

(a) an unstructured covariance matrix,
(b) a covariance matrix satisfying compound symmetry,
(c) a covariance matrix corresponding to a first-order autoregressive model for the repeat measures,
(d) a random coefficients model assuming a linear trend for each individual over the six post-randomization visits (see Crowder and Hand (1990) for details of the predicted form of the covariance matrix) and
(e) a random coefficients model assuming a quadratic trend for each individual over the six post-randomization visits.

Here none of the simpler structures provides an adequate fit for the observed covariance matrix. For the compound symmetry and first-order autoregressive models this is largely

TABLE 13
Initial model for the post-natal depression data†

Parameter	*Estimate*	*Standard error*	*z*	*P*
Constant	15.96	0.74	21.40	<0.0001
Treatment	1.84	0.52	3.52	0.0004
Linear time	−1.10	0.14	−7.68	<0.0001

†The treatment was coded 1 for placebo and −1 for active treatment.

TABLE 14
Parameter estimates for the final model for the post-natal depression data†

Parameter	*Estimate*	*Standard error*	*z*	*P*
Constant	8.80	3.61	2.43	0.01
Treatment	1.97	0.50	3.89	0.0001
Linear time	−1.99	0.54	−3.68	0.0002
Quadratic time	0.11	0.06	1.68	0.0922
Base-line	0.38	0.16	2.34	0.019

†This model can be written as

$$\mathrm{DEP} = 8.80 + 1.97\,\mathrm{GROUP} - 1.99X + 0.11X^2 + 0.38\,\mathrm{BASELINE}$$

where GROUP = 1 for placebo and GROUP = −1 for active treatment, X = 1, 2, 3, 4, 5, 6 for occasions 1–6 and BASELINE is the mean of the two pretreatment values. The estimate of the treatment difference is 2 × 1.97 = 3.94. The estimated standard error of the treatment difference is 2 × 0.50 = 1.00.

because the fitted values for the variances of the measurements on the six visits are constrained to be equal. The random coefficients models do allow these variances to differ but appear to fail here because of their inability to model the covariances adequately.

The fitted model is compared with the observed means in Fig. 10. The fit appears to be very good. The estimated difference between the two treatment groups is 3.94 with a 95% confidence interval of (1.94, 5.94). This difference remains the same for all post-treatment visits, which is a result that may have important implications for the design of future studies in this area.

Any statistical analysis should include a critical assessment of the modelling assumptions. Several diagnostics have been suggested for models used with repeated measures data including the *semivariogram* (Diggle, 1988) and Mahalanobis's D^2 (Gornbein *et al.*, 1992). Here a simple plot of the fitted mean response over the six post-treatment visits against the difference of the fitted and observed mean is shown in Fig. 11. There is no obvious pattern in this that might give cause for concern.

A simple analysis of these data by using the mean of the non-missing values over the six post-treatment visits as response and the mean of the two pretreatment values as covariate leads to the results shown in Table 16. Here the estimated treatment effect is 4.38 with 95% confidence interval (1.89, 6.87).

TABLE 15
Comparison of different covariance matrix structures for the post-natal depression data†

Structure	*Log-likelihood*	*No. of parameters*
Unstructured	−785.54	21
Compound symmetric	−834.88	2
Autoregressive	−825.45	2
Random coefficient (linear)	−823.51	4
Random coefficient (quadratic)	−810.51	7

†Testing twice the difference in the log-likelihoods as a χ^2-variable with degrees of freedom equal to the difference in the number of parameters indicates that none of the simpler structures can be chosen over the unstructured covariance matrix.

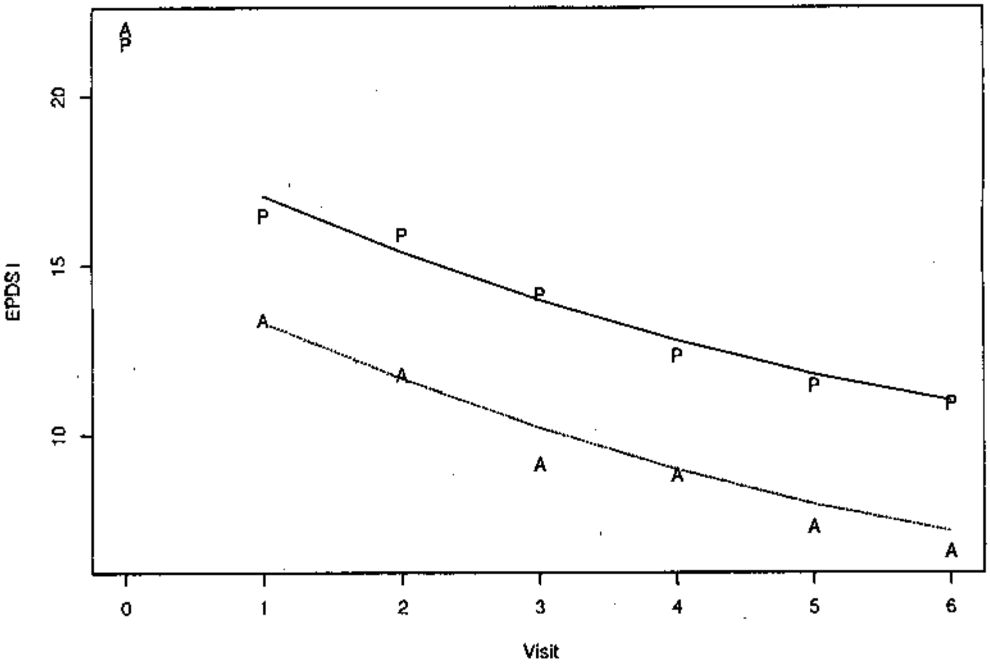

Fig. 10. Observed means and those predicted by the final model for the treatment trial of oestrogen: ———, placebo (P) group model; · · · · · ·, active (A) group model

5. Software for analysis of repeated measures

For the analysis of repeated measures at individual time points or when using the summary statistic approach with or without pre-randomization measures, almost any piece of current statistical software will be adequate. Again, to apply the analysis-of-variance procedures described in Section 3, a large number of packages would be suitable. For more involved analyses, e.g. those necessary when the data contain missing values and/or when a variety of models for the covariance structure are required, only relatively few packages are available. In this paper the BMDP program 5V has been used to obtain the results given for the oestrogen data. This program handles missing data by the maximum likelihood approach described in Little and Rubin (1987), and allows a number of structures for the covariance matrices of the repeated measures. Similar analyses are possible with SAS's procedure MIXED. It is also possible to analyse repeated measures data where the observations are not all made at the same set of time points with the program ML3. See Prosser *et al.* (1991) for details.

TABLE 16
Analysis of the treatment trial for post-natal depression by using the mean of six post-treatment measures as the response and the mean of pretreatment values as the covariate†

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*	*F*	*P*
Group	287.77	1	287.77	12.36	0.0009
Error	1350.65	58	23.29		

†The estimated regression coefficient is 0.49 and the adjusted means are 14.85 (placebo group) and 10.47 (active group). The estimated treatment difference is 14.85 − 10.47 = 4.38. The estimated variance of this difference is 23.29(1/27 + 1/34) = 1.55.

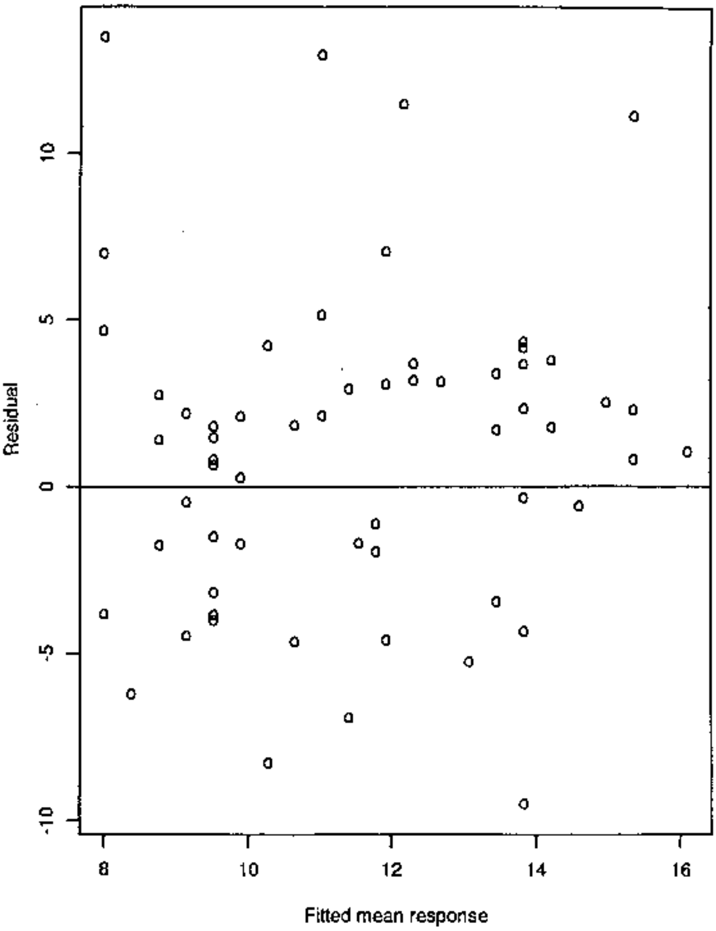

Fig. 11. Residuals of the fitted mean response for the oestrogen data

6. Summarizing remarks

Repeat measures data occur frequently in a variety of different disciplines. Methods of analysis range from a series of t-tests at each individual time point to the fitting of complex models that allow for missing data and a variety of covariance structures. Several researchers have made specific recommendations about the most appropriate techniques for dealing with this type of data. Ekstrom (1990), for example, suggests that the multivariate analysis-of-variance approach discussed in Section 3 is to be preferred in most situations, whereas Frison and Pocock (1992) argued that, in many medical investigations at least, the use of the average response over the repeated measures is often a sensible way to proceed. Perhaps the answer is that no single 'best method' is applicable to all cases simply because such data are collected in so many settings with such a variety of aims. In the examples discussed in this paper, for instance, simply analysing the mean summary measure for the cattle growth data would overlook important aspects of the data. For the treatment trial of oestrogen, however, a simple analysis of means leads to very similar conclusions and estimates as a more complex analysis. On balance, however, there are clear advantages for the busy applied statistician in the approach involving the calculation of a small number of *relevant* summary measures and comparing them across groups by using ordinary univariate methods. In particular this approach is easy both to explain and to perform. Faced with the difficult task of communicating results to the possibly statistically naïve researcher, such advantages should not be underestimated.

Acknowledgement

The comments made by a referee considerably improved this paper.

References

Broota, K. D. (1989) *Experimental Design in Behavioural Research.* New Delhi: Wiley Eastern.

Crowder, M. J. and Hand, D. J. (1990) *Analysis of Repeated Measures.* London: Chapman and Hall.

Davidson, M. L. (1972) Univariate versus multivariate tests in repeated measures experiments. *Psychol. Bull.*, **77**, 446–452.

Diggle, P. J. (1988) An approach to the analysis of repeated measures. *Biometrics*, **44**, 959–971.

Diggle, P. J., Liang, K. and Zeger, S. L. (1994) *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Ekstrom, D. (1990) Statistical analysis of repeated measures in psychiatric research. *Arch. Gen. Psych.*, **47**, 770–772.

Finney, D. J. (1990) Repeated measurements: what is measured and what repeats? *Statist. Med.*, **9**, 639–644.

Frison, L. and Pocock, S. J. (1992) Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design. *Statist. Med.*, **11**, 1685–1704.

Gornbein, J. A., Lazaro, C. G. and Little, R. J. A. (1992) Incomplete data in repeated measures analysis. *Statist. Meth. Med. Res.*, **1**, 275–295.

Greenhouse, S. W. and Geisser, S. (1959) On the methods in the analysis of profile data. *Psychometrika*, **24**, 95–112.

Hand, D. J. and Taylor, C. C. (1987) *Multivariate Analysis of Variance and Repeated Measures.* London: Chapman and Hall.

Heyting, A., Tolboom, J. T. B. M. and Essers, J. G. A. (1992) Statistical handling of drop-outs in longitudinal clinical trials. *Statist. Med.*, **11**, 2043–2061.

Huynh, H. and Feldt, L. S. (1976) Estimates of the Box correction for degrees of freedom for sample data in randomised block and split-plot designs. *J. Educ. Statist.*, **1**, 69–82.

Jones, M. C. and Rice, J. A. (1992) Displaying the important features of a large collection of similar curves. *Am. Statistn*, **46**, 140–145.

Kenward, M. G. (1987) A method for comparing profiles of repeated measurements. *Appl. Statist.*, **36**, 296–308.

Kumar, C., Gregoire, A. and Everitt, B. S. (1993) A treatment trial of oestrogen patches for post-natal depression. To be published.

Laird, N. M., Donnelly, C. and Ware, J. H. (1992) Longitudinal studies with continuous responses. *Statist. Meth. Med. Res.*, **1**, 225–247.

Lindsey, J. (1993) *Models for Repeated Measurements.* Oxford: Oxford University Press.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Matthews, J. N. S. (1993) A refinement to the analysis of serial data using summary measures. *Statist. Med.*, **12**, 27–37.

Matthews, J. N. S., Altman, D. G., Campbell, M. J. and Royston, P. (1990) Analysis of serial measurements in medical research. *Br. Med. J.*, **300**, 230–235.

Mauchly, J. W. (1940) Significance test for sphericity of a normal *n*-variate distribution. *Ann. Math. Statist.*, **29**, 204–209.

Pahwa, A. and Broota, K. D. (1981) Field-independence, field dependence as a determinant of colour-word interference. *Univ. Psychol. Res. J.*, **2**, 50–55.

Prosser, R., Rasbash, J. and Goldstein, H. (1991) *ML3—Software for Three-level Analysis.* London: Institute of Education.

Rouanet, H. and Lepine, D. (1970) Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *Br. J. Math. Statist. Psychol.*, **23**, 147–163.

Winer, B. J. (1971) *Statistical Principles in Experimental Design*, 2nd edn. New York: McGraw-Hill.