HOW TO ANALYSE YOUR REPEATED MEASURES DATA

DAVID J. HAND

INTRODUCTION

A fundamental assumption in many statistical models is that the observations are independent. However, in studies involving repeated measurement of the same variable on each experimental unit (on different occasions or under different conditions, for example) this assumption is unlikely to hold: observations on the same experimental unit are likely to be correlated. Data sets with this kind of structure are remarkably ubiquitous, arising in a broad range of disciplines. For this reason, especially in recent years, a great deal of research effort has been made in developing statistical techniques permitting the effective and valid analysis of such data. Moreover, because of the diversity of the different origins of such data, many different approaches to their analysis have been developed.

This presents the researcher with a problem of choice: which of the various methods are best suited to the data and research issues at hand?

It is this question which is the motivating force behind this review - Instead of a simple technical catalogue of the various methods, we have attempted to identify the important factors in determining appropriate choice of technique and then to describe methods of analysis with emphasis on these factors. For reasons of space we have restricted the discussion to the most popular methods and only to 'measured' on 'continuous' data. For similar reasons the descriptions are necessarily rather superficial. For more detail and numerical examples see Crowder and Hand (1990).

The body of this paper falls into two parts: first the discussion of the factors and second the descriptions of methods.

FACTORS INFLUENCING CHOICE OF METHOD

THE RESEARCH OBJECTIVE

It seems almost fatuous to state it, but choice of method is critically determined by the research question. Careful consideration must be given to

precisely what the researcher wants to know before a method can be selected. All too often, however, a method is chosen which addresses a similar but different research question. This may lead to loss of power in tests, to bias, and indeed, in worst cases, to incorrect conclusions. (This is not a problem unique to repeated measures analysis - see Hand (1991) for an example comparing two groups with only a single observation on each subject.)

Research questions can be positioned on a continuum ranging from general at one end to specific at the other. For example, in a repeated measures study involving comparing two groups we could ask general questions such as:

Q1: Do the group means change in different ways over time?

or we could move along the continuum becoming more specific:

Q2: Does group A get better faster than group B?

Q3: Is the linear trend over time in A greater than that in B?

Of course, many different specific questions can be asked. For example.

Q4: Are the linear trends equal?
Q5: Is the highest value reached by A greater than that reached by B?
Q6: Does A peak at the same time as B?

and so on.

Only the researcher can know which question is really to be addressed.

Questions such as these can be tackled by a general model building strategy in which an unconstrained model is compared to one constrained by the research question. For example, to address Q1 we would see how much better a fit to the data was provided by a model in which the means at each time were estimated separately within each group compared with a model in which the group means were constrained to follow the same pattern over time. Unfortunately, this does not completely determine the method. There are many possible models which might be fitted to the data. While fitting models will smooth away superfluous variation, yielding more powerful tests, to the extent that the model is incorrect bias will be introduced.

The questions used as illustrations above are examples of perhaps the more common kind of question, involving comparing the features of response curves in different groups. However, there are other types, and some of them are more

complex. For example, the researcher may wish to understand how relationships between variables (such as correlations) change over time. Even questions involving two groups may not be expressible as a comparison of features of the two curves. In a comparison of a group of extravert subjects with a group of introvert subjects galvanic skin response was measured hourly from 7 am until 11 pm. The question was not "Are the curves of means different?" (ie. is there an interaction?) but "Do the mean curves cross over?" The analysis also had to contend with some of the extra complexities to be discussed below since the numbers in the groups changed as time progressed because the subjects woke and went to sleep at different times.

DATA STRUCTURE

Data structure is a subset of the known properties of the data. For example, we may know that the experimental units were randomly allocated to groups. We may suspect a Markov dependence of measurement errors, but such a suspicion would arise from background theory and not from known properties of the data, not from the data structure. In repeated measures studies we can make the usual distinction between and within subjects structures. The former refers to the between experimental unit grouping structure and the latter to the structure of measurements on each unit. Examples of the latter include spacing over time (regular? the same for each unit?) and whether there are multiple measurements at each time. One distinction which can usefully be regarded as an aspect of data structure is between sequential treatment administration and situations where subjects are monitored without intervention. Laird and Ware (1982) call the former repeated measures studies and the latter growth curve studies. Covariate structure (once, at baseline? multiple times? etc) is also an aspect of data structure.

Missing values are a crucial aspect of data structure as far as choice of techniques goes - some methods of repeated measures analysis can only handle incomplete data in a clumsy and unsatisfactory way. Missing values can pose a serious problem with studies of this kind since the fact of multiple measurements often means that few subjects are complete. Rejecting incomplete subjects can lead to small sample size, not to mention the risk of bias. As a general principle, one should look particularly carefully at accounts of analyses undertaken with methods which cannot handle missing values - and ask oneself whether the data really were originally complete.

BACKGROUND THEORY

Background theory will lead to expectations regarding the shapes of the curves of the experimental units: should they asymptote to zero? will they increase over time? will they peak? If a method involving fitting individual curves to each unit is adopted then theory may suggest suitable families of curves.

Theory may also indicate what sort of error structure to fit. For example, sequential observations on humans (as in psychology) are unlikely to have compound symmetric structure.

COMPREHENSIBILITY

A statistical consultant's ultimate responsibility is to the client and as such it is important that the model, its assumptions, and its predictions should be comprehensible to the client. One implication is that the mathematically most sophisticated, "statistically optimal" technique may not be the practically best. A question statisticians should ask themselves when advising on choice of method is whether two alternatives (the "statistically best" and a simpler approximate method) are likely to yield essentially the same conclusions.

There is also a sociological issue of acceptability by journal editors. Comprehensibility sometimes, unfortunately, means doing it the way it has been done in the past. Of course, if the old way was inappropriate then persuasion is necessary.

ROBUSTNESS

Confidence in one's assumptions and the extent to which a method will yield incorrect conclusions should the assumptions be wrong will naturally influence choice of method.

PRACTICAL PERFORMANCE

Much statistical theory is based on asymptotic results and it may be that for small sample studies (often the case for repeated measures problems) the conclusions are not quite valid. Moreover there may be practical aspects beyond currently understood theory. A good example of this, though not in repeated measures, occurred in applying classical linear discriminant analysis to multivariate binary data. In this case different mean vectors necessarily implies

that the assumption of equal covariance matrices breaks down (except in certain special cases) so that theory tells us the method is inappropriate. And yet it can perform well in practice.

PARSIMONY

Increasing the complexity of a statistical model in general means that it will be able to fit the data more accurately. But, of course, it is not the data per se that we are trying to fit in inferential models - it is the mechanism generating the data. There is a risk, as complexity is increased, of overfitting, with consequent loss in power.

FLEXIBILITY

The chosen model, while being as parsimonious as possible, must also be as flexible as necessary. One could combine restrictions of comprehensibility and parsimony with information from background knowledge, data structure and other factors and adopt the simplest model yielding adequate flexibility. Alternatively, Diggle (1988) suggests using a single, general, flexible but parsimonious model. He bases his suggestions on the following criteria:

Specification of the mean profile must be sufficiently flexible to reflect
- (a) time trends within groups.
- (b) difference by trends between groups.

(ii) The specification of the error covariance matrix should be flexible but economical.

(iii) The method must be able to accommodate arbitrary patterns of occasions.

(iv) The method must be accompanied by diagnostics to assess goodness of fit.

SOFTWARE AVAILABILITY

In general, a statistical technique, no matter how appropriate, original, or powerful, will only be used if suitable software exists. While a consultant statistician may be able to use APL, S+, etc. to perform an unusual analysis, by far the vast majority of users will have to use one of the readily available packages. A brief review of what is available for repeated measures analysis in BMDP, SPSSX, and SAS is given in Crowder and Hand (1990).

A BRIEF REVIEW OF METHODS

BETWEEN GROUPS COMPARISONS AT EACH TIME

This method has very little to recommend it. The fact that many tests are being performed inflates the overall type I error rate, and moreover the inflation is in an unquantifiable way because of the dependencies between the tests. The method does not permit examination of patterns of change over time, which are often of central interest, and it also requires that observations are made at the same time on each experimental unit.

Sometimes this method is used to address the question "At what time do the groups begin to differ?" In fact this question probably does not have meaning as often as it is asked - differences often develop gradually. Where it is meaningful (for example, in detecting time of ovulation from body temperature measurements) other methods, such as the antedependence approach, are more suitable.

RESPONSE FEATURE ANALYSIS

Often researchers can identify particular features of the response curves which are of central interest to their research question.

Some examples of curve shapes are given in Figures 1 to 4 (taken, with permission, from Crowder and Hand, 1990).

Figure 1 (Figure 2.4 of Crowder and Hand) shows the body weights of rats measured on several occasions. The linear trend may be of interest. Figure 2 (Figure 5.1 of Crowder and Hand) shows growth curves of chicks. This demonstrates the typical fan shape obtained in such studies (and also shows the possibilities of bias since the lighter chicks seem to drop out of the study). Here a quadratic component might also be of interest. In Figure 3 (Figure 2.3 of Crowder and Hand), showing blood glucose levels at various times after a meal, peak, time to peak, or time to return to normal might be features of interest. Figure 4 (Figure 3.3 of Crowder and Hand) shows plasma ascorbic acid of patients on a particular diet. Here perhaps the time above some level would be a feature of interest.

The research question will, of course, determine what feature will be used, and this will often depend partly on background theory. For example, in pharmacokinetics the area under the curve (AUC) has a substantive meaning

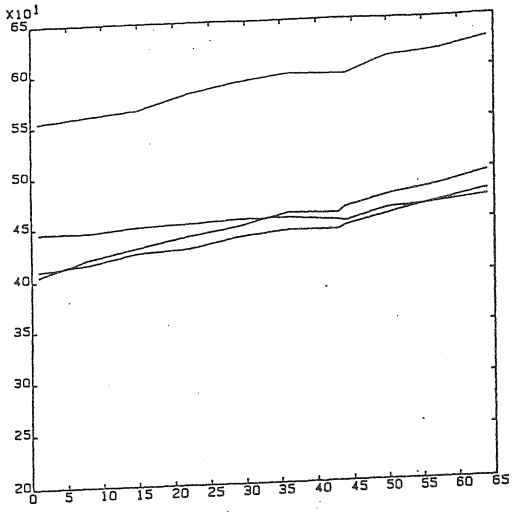

Figure 1: Body weights of rats over time

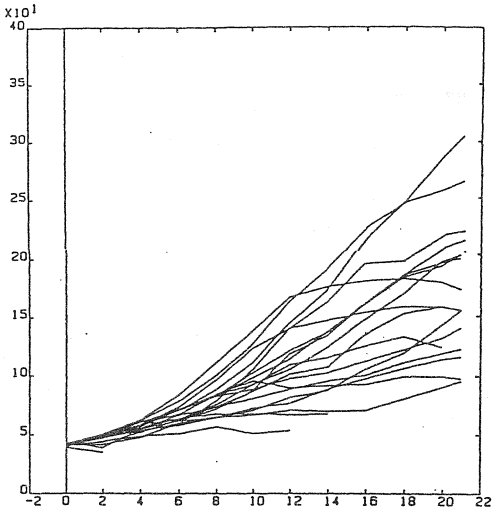

Figure 2: Growth curves of chicks

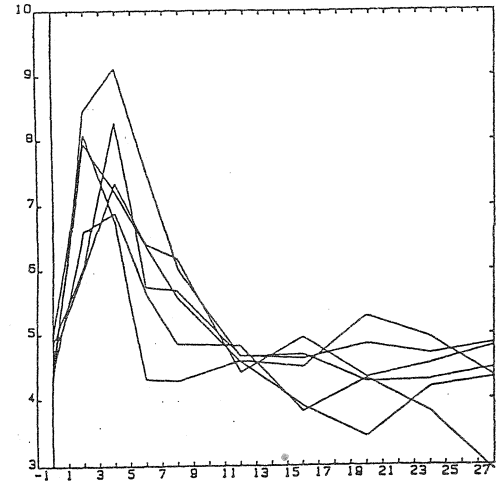

Figure 3: Blood glucose levels

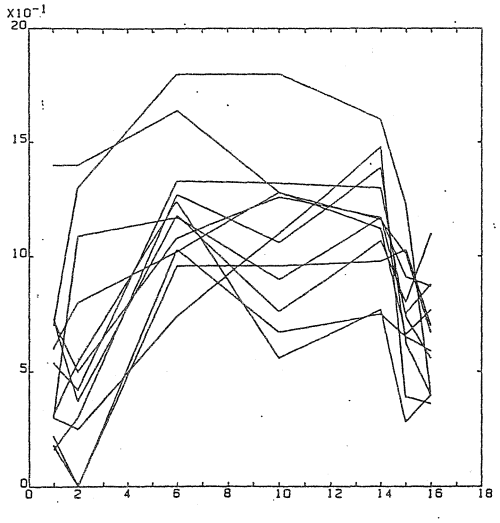

Figure 4: Plasma ascorbic acid

and in many areas growth rate or peak level has significance. However, it should be recognised that, even having identified the feature of interest, complications can arise. Choice of baseline can radically influence the AUC measure. One study the author was involved in had the later measurements spread very sparsely over a very long time period, by which time the level had returned almost to baseline, so that the curves had a very long right tail. Slight errors in measuring the later values had a substantial effect on the accuracy of the AUC feature.

This method does have a number of major advantages. It can handle different patterns of times for each unit including missing values, and it is also conceptually straightforward. Moreover, by reducing the multiple measurements on each individual to single feature scores, straightforward methods of analysis such as univariate analysis of variance of these single derived scores are possible. Software is thus readily available.

Although conceptually straightforward and easy to apply, the method may conceal latent problems. For example, with missing values, units with fewer observations will have their derived features estimated less accurately and ideally some account of this should be taken in the analysis. One can regard the random regression model, outlined below, as a more formal (conceptually less simple) extension of this method.

UNIVARIATE ANALYSIS OF VARIANCE

Response feature analysis combines all observations on each experimental unit into one (or more) scores summarising aspects of particular interest for each unit.

Unfortunately, not all researchers are able or willing to identify particular features in this way and instead want a more global analysis. The univariate analysis of variance approach does just this by regarding time as a factor in a mixed effects analysis of variance: subjects as a random effect and time as a fixed effect. This means that one can use standard packages to perform the analysis and this approach has been very popular in some disciplines (such as psychology). Unfortunately the method is not always valid. The F-tests are only valid if a condition termed "sphericity" holds.

If Σ is the error covariance matrix and P is a p x (p –1) matrix of orthonormal contrasts (there being p measurements on each unit) then Σ is said to satisfy sphericity if:

$$P' \Sigma P = \sigma^2 I$$

Alternatively, we require the ij th element of Σ, σ_{ij}, to satisfy $\sigma_{ij} = \alpha_i + \alpha_j + \lambda \delta_{ij}$ ($\lambda > 1$) with $\alpha_1, \ldots \alpha_p$, λ constants. Note that this means that if all variances are equal then all covariances are equal, a special case termed "compound symmetry" (which is sufficient, but not necessary, as is sometimes stated).

Sphericity is a rather special condition. It is, for example, equivalent to the condition that the variances of the differences between measurements at any time should be the same. This is unlikely to be true in situations involving repeated measurements over time where one might expect the variance of the difference between scores at nearby times to be smaller than the variance of the difference between scores at distant times. When the condition breaks down the type I error rate of the F-tests is inflated: too many true null hypotheses are rejected.

Fortunately a simple adjustment to the F-tests can be made to correct for non-sphericity. This is achieved by multiplying both numerator and denominator degrees of freedom by a measure of non-sphericity - a sphericity parameter, usually denoted by ε and defined as

$$\varepsilon = (\text{trace } P \Sigma)^2 / (p-1) \text{ trace } (P \Sigma)^2$$

Of course, to do this in practice ε must be estimated. One such estimate is obtained by plugging the usual maximum likelihood estimate for Σ in the above expression for ε, to yield the Greenhouse-Geisser estimate (Greenhouse & Geisser, 1959). Huynh & Feldt (1976), noting that this estimate was biased if $\varepsilon > 0.75$, suggested as an alternative

$$\tilde{\varepsilon} = \min [1, \{n (p-1) \hat{\varepsilon} - 2\} / (p-1) \{n - g - (p-1) \hat{\varepsilon}\}$$

where $\hat{\varepsilon}$ is the Greenhouse and Geisser estimate and g is the number of groups.

Non-sphericity may not always present problems. If there are only two measurements, of course, then the requirement vanishes (tests on patterns of change over time reduce to univariate tests) and this is true in general if a single contrast over measurements is being analysed. Sometimes more complex

analyses can be formulated as a sequence of univariate analyses on derived variables.

If there is a factorial structure to the within subjects design then derived variables can be produced, being contrasts of the original variables, which fall into natural groups which will be tested separately. In such a case sphericity will only be required within each of the groups.

The univariate analysis of variance method has the property that it can be applied if $p \geq n - g$, a property not shared by the multivariate approach outlined below. However, missing values cause a problem and it cannot be applied if the units have different time patterns.

Background theory may lead to ideas about the likely covariance structure and in this case one may be able to choose a more appropriate alternative technique.

MULTIVARIATE ANALYSIS OF VARIANCE

This approach has become more popular in recent years with the widespread availability of computer software such as MULTIVARIANCE and SPSSX-MANOVA. Here the p measurements on each unit are regarded as the components of a vector and between group comparisons are made using multivariate extensions of t-tests and analysis of variance. Particular aspects of the profiles over time can be studied by transforming the raw variables to yield suitable linear combinations.

The method is very flexible, making no restrictive assumptions about the covariance matrix Σ. However, unless p is small it is not parsimonious and will have low power. Similarly, since the tests involve inverting Σ it cannot be used if $p \geq n - g$ since then $\hat{\Sigma}$ will be singular.

Another problem is that whereas the F-test in univariate analysis of variance is the uniformly most powerful test invariant to rescaling there is no test in the multivariate case which is uniformly most powerful invariant to rescaling and rotation. Here the relative power of the tests depends on the type of departure from the null hypothesis. In general, however, any test based on $\lambda_1 \ldots \lambda_p$, the eigenvalues of $B\Sigma^{-1}$ (B being the between groups covariance matrix) will be invariant to rescaling and rotation. Four such tests are in common use and are typically given as computer output:

(i) Pillai - Bartlett trace $\Sigma \lambda_i / (1 + \lambda_i)$

(ii) Wilk's Lambda $\prod (1 + \lambda_i)^{-1}$

(iii) Hotelling Lawley trace $\Sigma \lambda_i$

(iv) Roy's largest eigenvalue statistic λ_1

These have complicated distributions, but may be approximated by χ^2 or F distributions.

The abundance of tests here leads to a question of choice. However, it seems that power is influenced by the "noncentrality structure" of the data: whether the group means lie in an approximately linear relationship or have a more diffuse distribution under the alternate hypothesis. In the former case (a "concentrated" noncentrality structure) power appears to decrease in order (iv), (iii), (ii), (i) while in the diffuse case it decreases as (i), (ii), (iii), (iv).

Of course, being able to choose between these structures requires a lot of background knowledge and one might instead like to rely on the suggestion that concentration must be extreme before it changes the ordering - hence (i) might be favoured. These issues are discussed further in Hand & Taylor (1987). The multivariate approach requires observations to be made at the same times for each unit (though the spacing can be irregular) and cannot handle missing values - the most common approach being to drop incomplete cases.

REGRESSION MODELS

The general multivariate analysis of variance model of section 3.4 can be written as:

$$\mathbf{X} = \mathbf{A}\,\xi\,\mathbf{D} + \mathbf{e}$$

Where $\mathbf{X}$ is the $n \times p$ data matrix, $\mathbf{A}$ is the between experimental units design matrix, ξ is the $g \times r$ parameter matrix, $\mathbf{D}$ is the $r \times p$ within units design matrix, and $\mathbf{e}$ is the $n \times p$ error matrix. Hypotheses, both between and within, on the values of the parameters can be specified in terms of matrices as H_0: $\mathbf{C}\,\xi\,\mathbf{M} = \mathbf{O}$.

Alternatively, letting y' be a row of X, we can write

$$= \mathbf{A}\,\xi\,\mathbf{D} + \mathbf{e}'$$

where here $\mathbf{A}$ is $1 \times p$ and $\mathbf{e}'$ is $1 \times p$. This can be expressed as $\mathbf{y} = \mathbf{D}' \mu + \mathbf{e}$ with $\mu = \xi' A'$ $(r \times 1)$.

We can now apply standard regression ideas and choose the parameters μ to minimise

$$\sum_{i=1}^{n} (y_i - D' \mu)' \Sigma^{-1} (y_i - D'\mu)$$

For multigroup problems we extend μ to be $rg \times 1$ (r parameters for each group).

We can also let $\mathbf{D}$ vary from subject to subject (replacing $\mathbf{D}$ by $\mathbf{D}_i$ in the above). This permits this approach to handle missing values (Σ needs to be replaced by Σ_i).

An important property of this approach is that Σ can be structured - for example in terms of a few parameters as $\Sigma = \Sigma(\emptyset)$. So, for example, the method includes the usual independence model ($\Sigma = \sigma^2 I$), compound symmetry

($\Sigma = \sigma^2 \mathbf{I}_p + \sigma_\alpha^2 J_p$), antedependence (observations at times i and j ($j > i$) are independent given the intervening $(j - i - 1)$ observations whenever $(j - i - 1) > s$ where s is a parameter to be chosen), and other Markov type models.

The possibility for structuring the covariance matrix leads us on to the next subsection which, in a sense, generalises the above.

RANDOM REGRESSION MODELS

An assumption which is often realistic is that the units in a group have the same basic shape but with the parameters specifying the curve varying from unit to unit. In particular, we might assume that each curve follows a straight line, but with its own slope and intercept parameters.

This idea leads to a model of the form

$$\mathbf{y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i$$

where β is a vector of fixed regression coefficients and $\mathbf{b}_i$ is a vector of random regression coefficients. Here we will assume $\mathbf{e}_i \sim N(\mathbf{O}, \mathbf{E}_i)$ and $\mathbf{b}_i \sim N(\mathbf{O}, \mathbf{B})$.

E_i must be constrained in some way, or else the model does not restrict Σ at all and a common constraint is to $E_i = \sigma^2 I$, the conditional independence approach.

Note that the above expression for y_i can alternatively be expressed as $y_i = X_i \beta + u_i$ with $u_i \sim N(O, Z_i \beta Z_i' + E_i)$, from which we see that the model is as in section 3.5 with $\sum_i = \sum(\emptyset_i)$ and $\emptyset_i = \{B, E_i\}$.

Such models are sometimes called two stage models because random variation occurs both within units (e_i) and between units (b_i).

Models of this kind are very attractive, permitting more extensive use to be made of background knowledge in choosing appropriate covariance machines. The idea that each unit is following its own curve, apart from random variation about that curve, and that each curve is sampled from a distribution is also appealing. The models are more parsimonious than methods using a full unstructured covariance matrix. At present they may not be so comprehensible to researchers, but this is due to lack of familiarity and this can be expected to change. In particular, the lack of familiarity has arisen from lack of accessible software to drive researchers to use this approach and this has now been remedied with BMDP 5V. (Of course, software which can be used by experts has long been available, but such programs do not open the technique to the wider community.) Such models can also easily handle missing values and irregular measurements patterns, perhaps differing between subjects.

CONCLUSION

A great deal of work has been done on repeated measures analysis in recent years, as is demonstrated by the extensive biblography in Crowder and Hand (1990).

Here we have only attempted to review the basis and there are many areas we have not touched upon, such as categorical data, nonlinear growth curves, and non-normal observations. However, it will be apparent from the methods that are described above that the range does permit a good match to be made between the problem and the technique. The factors outlined in section 2 are useful to consider when making such a match.

REFERENCES

Crowder M.J. and Hand D.J. (1990) *Analysis of repeated measures.* Chapman and Hall: London.

Greenhouse S.W. and Geisser S. (1959) On the methods in the analysis of profile data. Psychometrika, 24, p95 - 112.

Hand D.J. (1991) On comparing two treatments. In preparation.

Hand D.J. and Taylor C.C. (1987) *Multivariate analysis of variance and repeated measures,* Chapman and Hall: London.

Huynh H. and Feldt L.S. (1976) Estimation of the Box correction for degrees of freedom for sample data in randomised block and split-plot designs. *Journal of Educational Statistics,* 1, p69 - 82.

Laird N.M. and Ware J.H. (1982) Random effects models for longitudinal data. *Biometrics,* 38, p963 - 974.

Faculty of Mathematics

The Open University

Milton Keynes

MK7 6AA

England