Appl. Statist. (1994)
43, *No.* 1, *pp.* 49–93

Informative Drop-out in Longitudinal Data Analysis

By P. DIGGLE†

Lancaster University, UK

and M. G. KENWARD

University of Reading, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, June 16th, 1993,* Professor V. S. Isham *in the Chair*]

SUMMARY

A model is proposed for continuous longitudinal data with non-ignorable or informative drop-out (ID). The model combines a multivariate linear model for the underlying response with a logistic regression model for the drop-out process. The latter incorporates dependence of the probability of drop-out on unobserved, or missing, observations. Parameters in the model are estimated by using maximum likelihood (ML) and inferences drawn through conventional likelihood procedures. In particular, likelihood ratio tests can be used to assess the informativeness of the drop-out process through comparison of the full model with reduced models corresponding to random drop-out (RD) and completely random processes. A simulation study is used to assess the procedure in two settings: the comparison of time trends under a linear regression model with autocorrelated errors and the estimation of period means and treatment differences from a four-period four-treatment crossover trial. It is seen in both settings that, when data are generated under an ID process, the ML estimators from the ID model do not suffer from the bias that is present in the ordinary least squares and RD ML estimators. The approach is then applied to three examples. These derive from a milk protein trial involving three groups of cows, milk yield data from a study of mastitis in dairy cattle and data from a multicentre clinical trial on the study of depression. All three examples provide evidence of an underlying ID process, two with some strength. It is seen that the assumption of an ID rather than an RD process has practical implications for the interpretation of the data.

Keywords: Antedependence; Attrition; Conditional independence; Covariance structure; Drop-outs; Longitudinal data; Mastitis; Missing values; Repeated measurements; Serial correlation

1. Introduction

Longitudinal data consist of time sequences of measurements on many experimental or observational units. A convenient notation is $\{(y_{ij}, t_{ij}): j=1, \ldots, n_i; i=1, \ldots, m\}$, in which y_{ij} denotes the jth measurement on the ith of m units and t_{ij} the corresponding time at which the measurement is made. Typically, the primary objective of longitudinal data analysis is to describe the mean response as a function of time, treatment effects and possibly other covariates attached either to units or to individual measurements.

A convenient modelling framework for longitudinal data is the following general linear model with correlated errors. Let the random vector $\mathbf{Y}_i=(Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$ represent the sequence of measurements on the ith unit and $\mathbf{Y}=(\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_m^{\mathrm{T}})^{\mathrm{T}}$ the

† *Address for correspondence*: Department of Mathematics, Lancaster University, Lancaster, LA1 4YF, UK
E-mail: maa026@uk.ac.lancs.cent1

 0035–9254/94/43049

entire set of measurements on the m units. Define $\mathbf{t}_i$ and $\mathbf{t}$ similarly, and let $N=\Sigma_{i=1}^{m} n_i$. Then we assume that $\mathbf{Y}$ follows a multivariate Gaussian distribution,

$$\mathbf{Y} \sim \text{MVN}\{\mathbf{X}\boldsymbol{\theta},\ V(\mathbf{t},\ \boldsymbol{\phi})\}, \tag{1}$$

where $\mathbf{X}$ is an $N \times p$ matrix of explanatory variables and $V(\mathbf{t}, \boldsymbol{\phi})$ a block diagonal matrix with non-zero $n_i \times n_i$ blocks $V_i(\mathbf{t}, \boldsymbol{\phi})$. In model (1) the p-element parameter vector $\boldsymbol{\theta}$ affects the mean response and therefore is of direct interest, whereas the q-element parameter vector $\boldsymbol{\phi}$ is a set of nuisance parameters. The essential features of this modelling framework are the general linear model for the mean, the separate parameterizations of the mean and covariance structures, and the assumption that measurements from different units are uncorrelated. Likelihood-based inferences for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ then present no major computational difficulties, because the non-linear parts of the computations involve only the $n_i \times n_i$ matrices $V_i(\mathbf{t}, \boldsymbol{\phi})$, and the n_i are typically not large. This is in contrast with classical time series analysis where the data consist of one or a few long series of measurements.

Many researchers have proposed versions of model (1) using different specifications of covariance structure. See, for example, Laird and Ware (1982), Pantula and Pollack (1985), Ware (1985), Kenward (1987), Diggle (1988), Cullis and McGilchrist (1990), Verbyla and Cullis (1990), Jones and Ackerson (1990), Jones and Boadi-Boteng (1991) and Munoz *et al.* (1992).

A common phenomenon with longitudinal data is that of *drop-outs*, in which sequences of measurements on some units terminate prematurely; for example in animal experiments some animals may die during the course of the experiment. One important issue which then arises is whether the drop-out process is related to the measurement process. A useful classification of drop-out processes, following the terminology in Rubin (1976) and Little and Rubin (1987), is

(a) *completely random drop-out* (CRD)—the drop-out and measurement processes are independent;
(b) *random drop-out* (RD)—the drop-out process depends on the *observed* measurements, i.e. those preceding drop-out;
(c) *informative drop-out* (ID)—the drop-out process depends on the *unobserved* measurements, i.e. those that would have been observed if the unit had not dropped out.

Diggle (1989) and Ridout (1991) develop tests of the null hypothesis of CRD against an RD alternative. Under CRD, drop-outs are equivalent to randomly missing values and the data can be analysed straightforwardly by using model (1) which allows for the sequence lengths and times of measurement to differ between units.

Gould (1980) and the associated correspondence with Pledger and Hall (1982) contain an early discussion of the issues involved when the drop-out process is not CRD, but in the context of cross-sectional analyses of data from a clinical trial. Greenlees *et al.* (1982) and Glynn *et al.* (1986) consider the ID case. Again, this work is not in an explicitly longitudinal context although the issues raised are relevant to longitudinal data analysis. For longitudinal data, Laird (1988) gives an excellent discussion of how the drop-out process can affect the inferences about the measurement process. One of the most striking results is that the drop-out process can be ignored in the RD case provided that the required inferences concern the measurement process for notionally complete sequences, and those inferences are likelihood based. Murray

and Findlay (1988) illustrate the biases that can arise with non-likelihood-based methods of inference, such as analysis of complete sequences only, or of means calculated at each time point from those units that have not dropped out. Wu and Carroll (1988) consider ID in a random effects model, with each experimental unit following a linear time trend whose intercept and slope vary between individuals according to a bivariate Gaussian distribution. Wang *et al.* (1992) report on a simulation study to compare different methods of estimation under different assumptions about the drop-out process.

Our objective has been to develop a modelling strategy that accommodates CRD and RD as explicit special cases within an ID model. So far as we are aware, this is the first paper to propose a general model for ID in longitudinal measurement data and to develop associated methodology for likelihood-based inference. We emphasize that *drop-outs* are to be distinguished from *intermittent missing values*, in which an observed sequence has one or more gaps in it, or *unbalanced data*, in which the set of intended times of measurements is not common to all units.

Our particular approach has been coloured by our experience with a wide range of agricultural and biomedical applications in which the incidence of drop-outs has been sufficiently high to cast doubt on the results of conventional analyses which ignore the drop-out process. Others with different experiences may well view the problem somewhat differently, as for example in the social sciences (e.g. Heckman and Singer (1985)).

In Section 2 we define the model and argue that in general it is important to distinguish between all three of CRD, RD and ID. In Section 3 we develop likelihood-based methods of inference for our model. In Section 4 we report on a series of simulation experiments that address the following questions. Are inferences about the measurement process robust to misspecification of the drop-out process? How well can we estimate the parameters of the model from small samples? Our results, obtained from sample sizes of approximately 50 units, suggest that the methodology is quite adequate for the estimation of the parameters from the postulated drop-out model and can correct for the bias that appears in the RD analyses when the underlying process is ID. However, for the situations examined in the simulation study the biases seen in the CRD analyses, although consistent, were small and would not be of great practical concern. In Section 5 the methodology is applied to three real examples. All provide evidence for an ID process, and in two the evidence is very strong. Differences of practical importance between the ID and RD analyses are seen in the estimates of the mean responses.

2. General Model for Longitudinal Data with Drop-out

We develop our model for a single sequence of measurements on one unit. In view of the assumed independence between units, it is then easy to build this into a model for an entire set of longitudinal data, incorporating explanatory variables as appropriate. For the time being we ignore explanatory variables altogether. This is partly to avoid unnecessary complications in the notation, but also because we want to emphasize the structure of the stochastic components of the model, namely the measurements and the drop-outs. Let $\mathbf{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)^{\mathrm{T}}$ denote the generic n-element complete vector of measurements on an experimental unit and $\mathbf{t} = (t_1, \ldots, t_n)^{\mathrm{T}}$ the corresponding set of times at which the measurements are made. Let

$\mathbf{Y}=(Y_1, \ldots, Y_n)^{\mathrm{T}}$ denote the generic vector of observed measurements, with missing values coded as 0. The crucial assumption in our model for the drop-out process is that if an experimental unit is still in the study at time t_k its associated sequence of measurements $\{Y_j: j=1, \ldots, k\}$ follows the same joint distribution as that of the corresponding $\{Y_j^*: j=1, \ldots, k\}$. We therefore consider models for $\mathbf{Y}$ and $\mathbf{Y}^*$ which satisfy the relationship

$$Y_j = \begin{cases} Y_j^* & j=1, \ldots, D-1, \\ 0 & j \geqslant D, \end{cases} \tag{2}$$

where $2 \leqslant D \leqslant n$ identifies the drop-out time and $D=n+1$ identifies no drop-out.

Now, let $f^*(y; \boldsymbol{\theta}, \boldsymbol{\phi})$ denote the joint probability density function (PDF) of $\mathbf{Y}^*$ under the multivariate Gaussian model (1). Also, let $H_k=(y_1, \ldots, y_{k-1})$ denote an observed sequence of measurements up to time t_{k-1} and y_k^* the value that would be observed at time t_k if the unit did not drop out. Our model for the drop-out process allows the conditional probability of drop-out at time d to depend on the history of the measurement process up to and including time t_d, so that for $d \leqslant n$

$$P(D=d \mid \text{history}) = p_d(H_d, y_d^*; \boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters.

We now derive the joint distribution of the observed sequence $\mathbf{Y}$ via the sequence of conditional distributions for Y_k given $(Y_1, \ldots, Y_{k-1})=H_k$. Let $f_k^*(y \mid H_k^*; \boldsymbol{\theta}, \boldsymbol{\phi})$ denote the conditional univariate Gaussian PDF of Y_k^* given $(Y_1^*, \ldots, Y_{k-1}^*)=H_k^*$ and $f_k(y \mid H_k; \boldsymbol{\theta}, \boldsymbol{\phi})$ the conditional PDF of Y_k given $(Y_1, \ldots, Y_{k-1})=H_k$. Then it follows from equations (2) and (3) that

$$P(Y_k=0 \mid H_k, Y_{k-1}=0)=1, \tag{4}$$

$$P(Y_k=0 \mid H_k, Y_{k-1} \neq 0) = \int p_k(H_k, y; \boldsymbol{\beta}) f_k^*(y \mid H_k; \boldsymbol{\theta}, \boldsymbol{\phi})\, \mathrm{d}y \tag{5}$$

and, for $Y_k=y \neq 0$,

$$f_k(y \mid H_k; \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta}) = \{1-p_k(H_k, y; \boldsymbol{\beta})\} f_k^*(y \mid H_k; \boldsymbol{\theta}, \boldsymbol{\phi}). \tag{6}$$

Equations (4)–(6) determine the joint distribution of $\mathbf{Y}$. For a complete sequence $\mathbf{y}=(y_1, \ldots, y_n)$, and suppressing the dependence on the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$,

$$\begin{aligned} f(\mathbf{y}) &= f_1^*(y_1) \prod_{k=2}^{n} f_k(y_k \mid H_k) \\ &= f^*(\mathbf{y}) \prod_{k=2}^{n} \{1-p_k(H_k, y_k)\}, \end{aligned} \tag{7}$$

whereas for an incomplete sequence $\mathbf{Y}=(Y_1, \ldots, Y_{d-1}, 0, \ldots, 0)$ with drop-out at time t_d

$$\begin{aligned} f(\mathbf{y}) &= f_1^*(y_1) \left\{ \prod_{k=2}^{d-1} f_k(y_k \mid H_k) \right\} P(Y_d=0 \mid H_d) \\ &= f_{d-1}^*(\mathbf{y}^{d-1}) \left[\prod_{k=2}^{d-1} \{1-p_k(H_k, y_k)\} \right] P(Y_d=0 \mid H_d), \end{aligned} \tag{8}$$

where $\mathbf{y}^{(d-1)}=(y_1, \ldots, y_{d-1})^{\mathrm{T}}$, $f^*_{d-1}(\mathbf{y})$ denotes the joint PDF of the first $d-1$, non-zero, elements of $\mathbf{Y}^*$ and the product term within square brackets is absent if $d=2$.

Within this modelling framework, and following the terminology in Little and Rubin (1987), we distinguish three cases:

(a) ID—$p_k(\,)$ depends on y^*_k;
(b) RD—$p_k(\,)$ depends on H_k but not on y^*_k;
(c) CRD—$p_k(\,)$ depends neither on H_k nor on y^*_k.

In case (a), $p_k(\,)$ may also depend on H_k, but the crucial point is the dependence on y^*_k.

Both Diggle (1989) and Ridout (1991) used CRD as the null hypothesis. However, from some points of view the most important distinction is between RD and ID. To see why, note from equation (5) that under either CRD or RD

$$P(Y_k=0|H_k,\ Y_{k-1}\neq 0)=p_k(H_k;\ \boldsymbol{\beta}).$$

This, combined with equations (6)–(8) shows that the likelihood for the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ then separates into two components: one for $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and one for $\boldsymbol{\beta}$. This in turn implies that, under either CRD or RD, treating drop-outs simply as missing values leads to valid inferences about $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. However, these may not be the relevant inferences for the practical questions posed by the data. In particular, under CRD the unconditional mean structure of $\mathbf{Y}^*$ coincides with the conditional mean structure of $\mathbf{Y}$ given no drop-out, but this is not so under RD, except in the generally unrealistic case of uncorrelated Y_j. Clearly, the unconditional mean structure of $\mathbf{Y}$ differs from that of $\mathbf{Y}^*$ whatever assumptions are made about the drop-out process. In particular applications, any one of these three mean structures may be of primary interest.

For example, consider a time sequence of length $n=2$ in an RD model with $p_2(y_1)$ monotone decreasing in y_1. Then, the non-drop-outs at time 2 will contain predominantly units with large realized values of Y^*_1. If Y^*_1 and Y^*_2 are positively correlated, this predominance will transfer to Y^*_2, implying that $E(Y_2|\text{non-drop-out})>E(Y^*_2)$.

We conclude that, in general, it is important to distinguish between all three of CRD, RD and ID, and to develop valid inferences for $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ in all three cases. In the next section we develop likelihood-based inferences for particular parametric models within the general framework of this section.

3. Particular Models and Likelihood Inference

The results in Section (2) show how to construct a likelihood for $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ from parametric specifications of the measurement process and the drop-out process. We now examine each of these in turn.

3.1. *Covariance Structure of Measurement Process*

With regard to the measurement process, under the general linear model (1), the joint PDF $f^*_k(\mathbf{y}^{(k)})$ of the first k non-missing measurements and the conditional PDF $f^*_k(y_k|H_k)$ of the kth non-missing measurement given the preceding $k-1$ measurements follow from standard results on the multivariate Gaussian distribution. See, for example, chapter 6 of Chatfield and Collins (1980). Let $\boldsymbol{\mu}=(\mu_1, \ldots, \mu_n)^{\mathrm{T}}$

denote the mean response vector for a complete sequence of n measurements $\mathbf{Y}^* = (Y_1, \ldots, Y_n)^T$ on a single unit and $\mathbf{y}$ the realization of $\mathbf{Y}^*$. Note that $\boldsymbol{\mu} \equiv \boldsymbol{\mu}(\boldsymbol{\theta})$. Write $\boldsymbol{\mu}^{(k)}$ for the first k elements of $\boldsymbol{\mu}$. Let $\mathbf{V}$ denote the covariance matrix of $\mathbf{Y}^*$. Note that $\mathbf{V} \equiv V(\boldsymbol{\phi})$. Write $\mathbf{V}^{(k)}$ for the leading $k \times k$ submatrix of $\mathbf{V}$, $\mathbf{c}^{(k)} = (c_1^{(k)}, \ldots, c_k^{(k)})$ for the k-element vector of covariances,

$$c_j^{(k)} = \text{cov}(Y_j^*, Y_{k+1}^*), \qquad j = 1, \ldots, k,$$

and $v_{kk} = \text{var}(Y_k^*)$. Then, for each of $k = 1, \ldots, n-1$, $f_k^*(\mathbf{y})$ is k variate Gaussian with mean vector $\boldsymbol{\mu}^{(k)}$ and covariance matrix $\mathbf{V}^{(k)}$. Also, $f_{k+1}^*(y_{k+1} | H_{k+1})$ is univariate Gaussian with mean

$$m_{k+1} = \mu_{k+1} + \mathbf{c}^{(k)T}(\mathbf{V}^{(k)})^{-1}(\mathbf{y}^{(k)} - \boldsymbol{\mu}^{(k)}) \tag{9}$$

and variance

$$w_{k+1} = v_{k+1,k+1} - \mathbf{c}^{(k)T}(\mathbf{V}^{(k)})^{-1}\mathbf{c}^{(k)}. \tag{10}$$

For efficient updating of equations (9) and (10) we note that

$$\mathbf{V}^{(k+1)} = \begin{pmatrix} \mathbf{V}^{(k)} & \mathbf{c}^{(k)} \\ \mathbf{c}^{(k)T} & v_{k+1,k+1} \end{pmatrix}$$

and use the result that

$$\begin{pmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{x}^T & b \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C} & \mathbf{y} \\ \mathbf{y}^T & e \end{pmatrix} \tag{11}$$

where

$$e = (b - \mathbf{x}^T\mathbf{A}^{-1}\mathbf{x})^{-1}, \tag{12}$$

$$\mathbf{y} = -e\mathbf{A}^{-1}\mathbf{x} \tag{13}$$

and

$$\mathbf{C} = \mathbf{A}^{-1} + d(\mathbf{A}^{-1}\mathbf{x})(\mathbf{A}^{-1}\mathbf{x})^T. \tag{14}$$

Various parametric forms for $\mathbf{V}$ have been proposed in the references cited earlier. In many applications, it is sensible to think of $\mathbf{Y}^*$ as arising from three random components: a zero-mean random variable U representing the intrinsic level of response of the unit in question, a serially correlated stochastic process $\{W(t): t \in \mathscr{R}\}$ and an uncorrelated sequence Z_j, $j = 1, \ldots, n$, representing measurement errors or sampling variation. Then

$$Y_j^* = U + W(t_j) + Z_j, \qquad j = 1, \ldots, n, \tag{15}$$

and

$$V = \nu^2\mathbf{J} + \sigma^2\mathbf{G} + \tau^2\mathbf{I}, \tag{16}$$

where $\mathbf{J}$ is an $n \times n$ matrix with all its elements equal to 1, $\mathbf{G}$ is an $n \times n$ matrix with elements

$$g_{jk} = \text{corr}\{W(t_j), W(t_k)\}$$

and $\mathbf{I}$ is the $n \times n$ identity matrix. Diggle (1988) used this model with

$$g_{jk} = \exp\{-\alpha(t_j - t_k)^h\}, \tag{17}$$

for $h=1$ or $h=2$. Under this model, $\boldsymbol{\phi} = (\nu^2, \sigma^2, \tau^2, \alpha)$. In what follows we shall assume that the model for the covariance structure of $\mathbf{Y}^*$ is given by equations (16) and (17) with $h=1$. However, we emphasize that the ideas are equally applicable to other parametric models and we give one such example, using the antedependence covariance structure (Gabriel, 1961; Kenward, 1987) in Section 5.

3.2. *Drop-out Process*

We now consider the drop-out process and the specification of the conditional probability $p_k(H_k, y; \boldsymbol{\beta})$ in equation (3). In the spirit of regression modelling, we propose a logistic linear model,

$$\text{logit}\{p_k(H_k, y; \boldsymbol{\beta})\} = \beta_0 + \beta_1 y + \sum_{j=2}^{k} \beta_j y_{k+1-j}. \tag{18}$$

A possible extension of model (18) would be to allow $\boldsymbol{\beta}$ to depend on external covariates, some of which may vary with time. We return to this point shortly. One motivation for model (18) is that we might expect the probability of drop-out at time t to depend on a discounted integral,

$$\int_0^t \omega(s)\, y(t-s)\, \mathrm{d}s, \tag{19}$$

of the history of the measurement process up to and including time t, where $\omega(\,)$ is a weight function which determines the discounting over time. Then, the expression $\beta_1 y + \Sigma_{j=2}^{k} \beta_k y_{k+1-j}$ can be interpreted as a quadrature formula for integral (19), in which the weight function $\omega(\,)$ is chosen empirically to give the best fit to the data. Under this interpretation, it would be natural to assume a monotone decreasing weight function which, at least for equally spaced times of measurement t_j, would imply a monotone decreasing sequence β_k. This in turn suggests that, as soon as we reject CRD and allow any form of dependence between the drop-out process and the measurement process, RD becomes implausible by comparison with ID. However, this argument is somewhat simplistic. Firstly, 'drop-out at t_d' means 'drop-out somewhere between t_{d-1} and t_d'. Second, the propensity to drop out may derive from characteristics of the unit that are constant over time, e.g. the realized value of the random effect U in equation (15). A further consideration is that drop-out may depend on external factors, possibly related to the measurement process, rather than on the measurement process itself. For example, in a medical study a patient may fail to keep a return appointment because of their general state of health; in an animal experiment, an animal may die as a delayed response to undue stress early in the experiment. Or it may be that the drop-out probabilities simply change with time; for example, drop-out may be more prevalent in the later stages of a study. These considerations motivate extending model (18) by making β_0 a function of covariates w_{qk} at time t_k, e.g. a linear function

$$\beta_0 = \sum_{q=1}^{r} \beta_{q0} w_{qk}. \tag{20}$$

Note that the same covariates can appear both in the model for the drop-out probabilities and in the model for the mean profiles $\mathbf{X}\boldsymbol{\theta}$.

We summarize the above discussion as follows.

(a) The true nature of the drop-out process is likely to be complex.
(b) The logistic regression formulation (18) is a reasonable empirical model.
(c) The extension (20) permits investigation of possible relationships between the drop-out process and covariates, including time.

In what follows we assume model (18). The extension to model (20) is straightforward.

3.3. *Likelihood Function*

We now assemble these results into an explicit expression for the likelihood function. For m units, let $\mathbf{y}_i=\{y_{ij}: j=1, \ldots, d_i-1\}$ denote the observed measurements on the ith unit, with d_i indicating the drop-out time if $2\leqslant d_i\leqslant n$ and $d_i=n+1$ identifying no drop-out. Let $f_i^*(\mathbf{y}_i)$ denote the joint PDF of the d_i-1 available measurements from the ith unit,

$$\log f_i^*(\mathbf{y}_i)=-\{(d_i-1)/2\}\log(2\pi)-\tfrac{1}{2}\log|\mathbf{V}^{(d_i-1)}(\boldsymbol{\phi})|-\tfrac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}^{(i)})^{\mathrm{T}}\mathbf{V}^{(d_i-1)}(\boldsymbol{\phi})^{-1}(\mathbf{y}_i-\boldsymbol{\mu}^{(i)}), \tag{21}$$

where $\boldsymbol{\mu}^{(i)}$ represents the relevant d_i-1 elements of $\mathbf{X}\boldsymbol{\theta}$. From model (18),

$$\log\{1-p_k(H_k, y_k)\}=-\log\left\{1+\exp\left(\beta_0+\sum_{j=1}^{k}\beta_j y_{k+1-j}\right)\right\}. \tag{22}$$

Then, the log-likelihood for $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ based on data $\{\mathbf{y}_i: i=1, \ldots, m\}$ is

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta})=L_1(\boldsymbol{\theta}, \boldsymbol{\phi})+L_2(\boldsymbol{\beta})+L_3(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta})$$

where

$$L_1(\boldsymbol{\theta}, \boldsymbol{\phi})=\sum_{i=1}^{m}\log f_i^*(\mathbf{y}_i) \tag{23}$$

$$L_2(\boldsymbol{\beta})=\sum_{i=1}^{m}\sum_{k=2}^{d-1}\log\{1-p_k(H_{ik}, y_{ik})\} \tag{24}$$

and

$$L_3(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta})=\sum_{i:d_i\leqslant n}\log P(D=d_i|\mathbf{y}_i). \tag{25}$$

The explicit forms of $L_1(\,)$ and $L_2(\,)$ require substitution from equations (21) and (22) respectively. For $L_3(\,)$, we use equation (5) for $P(D=d_i|\mathbf{y}_i)$ in conjunction with equations (9), (10), (18) and a numerical integration rule. Note that the required inverses of the matrices $\mathbf{V}^{(k)}$ for $k=1, 2, \ldots, n$ are obtained efficiently by using the updating formulae (11)–(14). For maximization of the overall log-likelihood $L(\,)$ we have used the simplex algorithm of Nelder and Mead (1965).

The logistic form for the drop-out probabilities also allows us to approximate the integral in equation (5). Using the probit approximation to the logit transformation

we can apply the result described in Zeger *et al.* (1988) to obtain, omitting the unit index i,

$$\text{logit}\,[E^*\{P(D=d\,|\,\mathbf{y}^{(d-1)})\}] \simeq \beta_0^* + \sum_{j=2}^{d} y_{d-j+1}\beta_j^* \tag{26}$$

where the expectation E^* is taken over the conditional distribution of the missing observation Y_d^*,

$$\beta_0^* = a\,[\beta_0 + \beta_1\{\mu_d - \boldsymbol{\mu}^{(d-1)\mathrm{T}}(\mathbf{V}^{(d-1)})^{-1}\mathbf{c}^{(d-1)}\}] \tag{27}$$

and

$$(\beta_d^*, \ldots, \beta_2^*)^{\mathrm{T}} = a\{(\beta_d, \ldots, \beta_2)^{\mathrm{T}} + \beta_1\mathbf{c}^{(d-1)\mathrm{T}}(\mathbf{V}^{(d-1)})^{-1}\} \tag{28}$$

for $a^{-2} = 1 + \beta_1^2 c^2 w_d$ and $c = (16\sqrt{3})/15\pi \simeq 0.588$. The components w_d, $\boldsymbol{\mu}^{(d)}$, $\mathbf{V}^{(d)}$ and $\mathbf{c}^{(d)}$ are defined in equations (9) and (10). Note that Wu and Carroll (1988) use a probit model to relate the probability of drop-out to the random intercept and slope in their random effects model.

Finally, if RD holds, then equation (25) simplifies through

$$\log P(D=d_i\,|\,\mathbf{y}_i) = \beta_0 + \sum_{j=1}^{d_i}\beta_j y_{i,d_i+1-j} - \log\left\{1 + \exp\left(\beta_0 + \sum_{j=1}^{d_i}\beta_j y_{i,d_i+1-j}\right)\right\}.$$

$L_3(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta})$ then depends only on $\boldsymbol{\beta}$, and the likelihoods for $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and for $\boldsymbol{\beta}$ can be maximized separately. In any event, we recommend separate maximization of $L_1(\,)$ and of $L_2(\,)+L_3(\,)$ assuming RD as a way of obtaining initial values for the full maximization of $L(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta})$. Although only $L_1(\,)$ is required to make valid inferences about the marginal process $\mathbf{Y}^*$ when RD holds, the full likelihood is still needed for inferences about the conditional process $\mathbf{Y}^*\,|\,$non-drop-out.

We have assumed throughout this development that no parameters are common to both $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $\boldsymbol{\beta}$. Although this will often be a reasonable assumption, when it does not hold, separate maximization of $L_1(\,)$ would still give consistent, although not necessarily fully efficient, estimators for the marginal process $\mathbf{Y}^*$. This answers a conjecture by Shih (1992).

4. Simulation Study

In this section we investigate the behaviour of the proposed likelihood analysis through a simulation study. This also gives us the opportunity to make a comparison with the simpler RD analysis for data generated under an ID process, i.e. we compare the behaviour of the likelihood component $L_1(\boldsymbol{\theta}, \boldsymbol{\phi})$ in equation (23) with that of the full likelihood $L_1(\boldsymbol{\theta}, \boldsymbol{\phi})+L_2(\boldsymbol{\beta})+L_3(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta})$. We also include for comparison the behaviour of ordinary least squares (OLS) estimators.

We need to specify three aspects of the models to be simulated: the linear model structure, the covariance structure and the drop-out model. The linear model structure is central as it is the bias in the estimation of these parameters that represents the main issue. We therefore base our simulation study on two structures for the mean profiles of the repeated measurements: a comparison of two simple linear regressions and a four-period crossover trial. We also want to assess the behaviour of the process in moderate sample sizes, so for both set-ups we use approximately 50 units.

The comparison of trends over time frequently arises in the analysis of longitudinal data. The simple linear regression set-up allows us to investigate the behaviour of such estimates. Two groups are used, with 25 units assigned to each group, and 10 equally spaced repeated measurements. In the first group the mean level is set constant over time; in the second it declines at a constant rate of -1. Both lines start from the same intercept of 10. Straight lines allowing different slopes and intercepts are fitted to each group and the effect of interest is the difference in slopes. This can be summarized as follows. For the variable Y_{ij}^* at time j in group i:

$$E(Y_{ij}^*) = \gamma_i + \eta_i t_j, \qquad i = 1, 2,$$

where, for the simulation, $\gamma_1 = \gamma_2 = 10$, $\eta_1 = 0$ and $\eta_2 = -1$. We are interested in the difference in slopes, $\delta = \eta_1 - \eta_2$.

A stationary first-order autoregressive process was used to generate the residual component of the repeated measurements, implying a covariance structure of the form

$$V = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^9 \\ \rho & 1 & \rho & \ldots & \rho^8 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^9 & \rho^8 & \ldots & \rho & 1 \end{pmatrix},$$

a special case of the general covariance structure in equation (16) with $\nu^2 = \tau^2 = 0$. Arguably, this is too simple to be practically realistic but it serves the purpose of introducing within-unit correlation in a convenient and parsimonious way. In the simulations, σ^2 was set equal to 1, and ρ took the values 0.5, 0.75 and 0.9.

For each value of ρ, 1000 sets of data were generated by using Gaussian pseudorandom variates. For each set, CRD, RD and ID processes were applied according to the logistic model (18):

$$\text{logit}\{p_k(H_k, y_k; \boldsymbol{\beta})\} = \beta_0 + \beta_1 y_k + \beta_2 y_{k-1}$$

with

$$\begin{aligned} \beta_1 = \beta_2 = 0 \quad & \text{(CRD)}, \\ \beta_1 = 0 \quad & \text{(RD)}, \\ \beta_2 = 0 \quad & \text{(ID)}. \end{aligned}$$

Note that the ID process assumes no direct dependence on the previous observation. This is an extreme case of the ID model, which we chose because we expected that results from simulations with the dependence distributed between both current and previous observations would be intermediate between the results from the RD and ID assumptions defined above. Some preliminary trials with non-zero values for both β_1 and β_2 confirmed this.

In the simulations, β_0 was set to a small value so that the drop-out rate was very low in the CRD case. Values of β_1 and β_2 were chosen to produce proportions of missing values equal to approximately 33%, 50% and 66%. For a given value of $\boldsymbol{\beta}$ the proportion of values that are missing fluctuates slightly as the covariance parameters are altered; hence in the tables of results a range of percentages is associated with each setting of $\boldsymbol{\beta}$. These correspond to the smallest and the largest of the average number of missing values from each collection of 1000 simulations.

TABLE 1
Percentage bias in the estimate of slope difference from the simulation study of the regression model

Average % missing	*Drop-out mechanism*	*Relative bias (%) for the following values of ρ:* 0.5			0.75			0.9		
		Estimator			*Estimator*			*Estimator*		
		OLS	*RD*	*ID*	*OLS*	*RD*	*ID*	*OLS*	*RD*	*ID*
0	CRD	0	0	0	0	0	0	0	0	0
27–31	RD	−1	0	0	−3	0	0	−10	0	0
	ID	−1	−1	1	−4	−2	0	−12	−2	0
44–50	RD	−3	1	1	−8	1	1	−17	1	1
	ID	−5	−4	1	−10	−5	2	−22	−6	1
58–67	RD	−6	2	4	−17	2	3	−29	2	2
	ID	−17	−14	4	−25	−15	4	−37	−10	3

For each set of simulated data the RD and ID models were fitted by using maximum likelihood as developed in Sections 2 and 3. The *relative* bias in δ is presented in Table 1 as a percentage rounded to the nearest integer. There is negligible bias in the likelihood-based estimators when the data are generated under the RD process. Both models are appropriate in this case, although the ID model includes a redundant parameter, β_1. Some negative bias is apparent when the RD model is fitted to data generated under an ID process and, as should be expected, this increases in absolute size with an increasing proportion of drop-outs. This bias is nevertheless considerably smaller than that in the OLS estimators, which implies that the RD likelihood analysis is providing some protection against the bias induced by the ID process. The increase in OLS bias with increasing ρ implies that this protection is more effective with greater within-unit correlation. This is to be expected from the form of the approximate relationship between the RD and ID likelihoods defined in equations (26)–(28). The difference between the likelihoods decreases as the conditional variance of Y_j^* given Y_{j-1}^* decreases.

The small positive bias in the other likelihood-based estimators that appears under the higher drop-out rates is bias associated with the smaller implicit sample sizes. For example, on average only a third of the complete set of observations contribute to the estimator under the highest drop-out rate. This bias disappears when the sample size is increased.

We turn now to the crossover design. For this we use a four-period four-treatment Williams square arrangement with 12 units assigned to each of four sequences of treatments (Williams, 1949):

sequence 1	A	B	C	D
sequence 2	B	D	A	C
sequence 3	C	A	D	B
sequence 4	D	C	B	A

This Williams square is a particular example of a Latin square in which each treatment follows each other except itself equally often. This provides the estimates of the treatment effects with some robustness against certain forms of carry-over. For the

simulations we use a model in which there are no carry-over effects. For Y_{ij}^* corresponding to period j in sequence i we have the model

$$E(Y_{ij}^*) = \gamma_j + \alpha_{l[i,j]},$$

where γ_j is the jth period mean, α_l is the effect associated with treatment l and $l[i,j]$ is the treatment occurring in period j of sequence i. The use of the crossover arrangement allows us to consider in the simulation study both the estimation of effects associated with time varying covariates (the treatment differences), albeit in a rather balanced arrangement, and the estimation of means at particular times, the γ_j. From the symmetry of the design we can consider any treatment difference, and in the following we present the results from a single such difference $\alpha_A - \alpha_B$ with a true value of -5. Given the occurrence of drop-out, the period means do not have the same symmetry. We present the bias in the estimator of the last of these, γ_4. The bias in each of the other period means is consistently less than this. For the simulated data we have $\gamma_j = 10$, $j = 1, \ldots, 4$. In the simulations a simple uniform covariance structure was used for the repeated measurements:

$$V = \tau^2 \mathbf{I} + v^2 \mathbf{J},$$

with the within-unit variance τ^2 set equal to 1 and the between-unit variance component v^2 set equal to 0.001, 0.82 and 9.0, corresponding to within-unit correlations of 0.00, 0.45 and 0.90 respectively. This is again a special case of the general covariance structure in equation (16).

The drop-out mechanisms and overall structure of the simulations follow the same pattern as for the regression set-up above. Given the fewer repeated measurements, however, the proportions of missing data were chosen to be smaller, approximately 16%, 33% and 50%.

Tables 2 and 3 contain the observed biases from the simulations under the crossover model. The relative bias in the estimate of the fourth period mean, γ_4, is presented in Table 2, and the relative bias in the treatment difference, $\alpha_A - \alpha_B$, in Table 3.

Both sets of results have the same broad pattern as those observed for the regression analysis, although the absolute sizes of the biases are smaller. From the RD model fitted to data generated under the ID process the estimate of the period mean has

TABLE 2
Percentage bias in the estimate of the period 4 mean from the simulation study of the crossover model

Average % missing	*Drop-out mechanism*	*Relative bias (%) for the following within-unit correlations:* 0.00			0.45			0.90		
		Estimator			*Estimator*			*Estimator*		
		OLS	*RD*	*ID*	*OLS*	*RD*	*ID*	*OLS*	*RD*	*ID*
0	CRD	0	0	0	0	0	0	0	0	0
14–16	RD	0	0	0	−1	0	0	−11	0	0
	ID	0	0	−1	−2	0	−1	−12	−1	0
32–35	RD	0	0	0	−3	0	0	−20	0	0
	ID	−1	−1	0	−4	−3	0	−22	−3	0
45–50	RD	0	0	0	−5	0	0	−30	0	0
	ID	−3	−3	0	−8	−5	0	−33	−4	0

TABLE 3
Percentage bias in the estimate of a treatment difference from the simulation study of the crossover model

Average % missing	*Drop-out mechanism*	*Relative bias (%) for the following within-unit correlations: 0.00*			*0.45*			*0.90*		
		Fitted model			*Fitted model*			*Fitted model*		
		OLS	*RD*	*ID*	*OLS*	*RD*	*ID*	*OLS*	*RD*	*ID*
0	CRD	0	0	0	0	0	0	0	0	0
14–16	RD	0	0	0	0	0	0	−3	0	0
	ID	0	1	0	1	1	0	4	1	0
32–35	RD	0	0	0	1	0	0	−3	0	0
	ID	2	2	0	2	2	0	9	0	0
45–50	RD	0	0	0	1	0	0	−3	0	0
	ID	4	3	0	5	4	0	12	3	0

a negative bias and that of the treatment difference a positive bias. We conjecture that the opposite signs of the two biases arise from the non-orthogonality between treatments and periods that is induced by the missing data. The nature of the ID in the simulations causes downward bias in the period effects. Hence, correction of the treatment effects for period effects under non-orthogonality reverses this bias in the treatment effects. The absolute size of the bias increases with increasing proportion of drop-outs. Unlike the regression analysis, however, the bias under other combinations of model and drop-out process is negligible, even with the high proportions of missing values. Again, the OLS estimator is appreciably more biased than the RD estimator when there is high within-subject correlation. From a practical point of view the observed bias when using the RD analysis under the ID process is small.

5. Examples

5.1. *Milk Protein Trial*

The data for the milk protein trial are taken from Verbyla and Cullis (1990). They consist of assayed protein content of milk samples taken weekly from each of 79 cows. The cows were randomly allocated to one of three diets: barley, mixed barley–lupins and lupins, with 25, 27 and 27 animals in the three groups. Measurements were taken for up to 19 weeks, but there were 38 (48%) drop-outs from week 15 onwards, corresponding to cows who stopped producing milk before the end of the experiment. The primary objective of the experiment was to describe the effects of diet on the mean response profile over time. Our particular interest here is to ask how our interpretation of these effects might be influenced by the drop-out process. Fig. 1 shows the three observed mean response profiles for the data; note that from week 15 onwards the means are calculated only from cows still producing milk.

Previous analyses of these data are reported by Diggle (1990), chapter 5, and Verbyla and Cullis (1990). Diggle analyses the entire data set under the implicit assumption of RD and concludes that the mean response profiles can be described by a model of the form

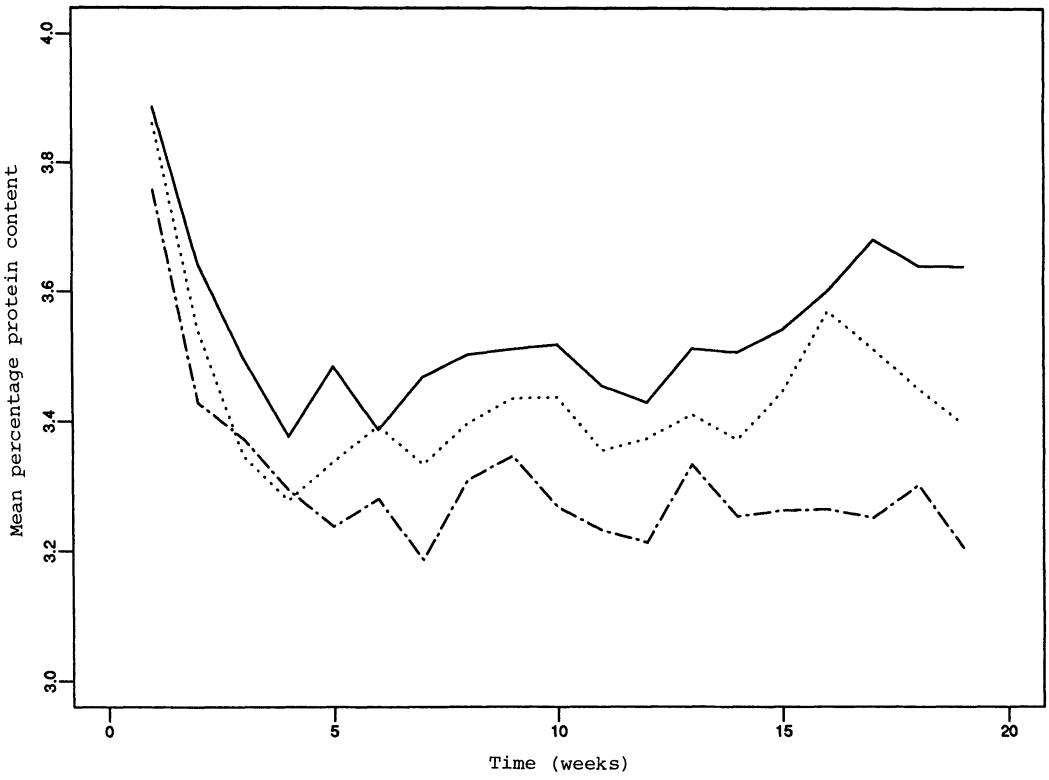

Fig. 1. Observed mean response profiles for the milk protein data: ——, barley diet; · · · · ·, mixed diet; —·—·, lupins diet

$$\mu_{ij} = \begin{cases} \gamma_i - j\alpha & j \leqslant 3, \\ \gamma_i - 3\alpha & j > 3 \end{cases} \tag{29}$$

where $i = 1, 2, 3$ denotes diet and time j is measured in weeks. Verbyla and Cullis ignore the data from the first three weeks, use 16 parameters to describe the mean response μ_{1j}, $j = 1, \ldots, 16$, for diet 1 in each of the last 16 weeks and describe the contrasts $\mu_{2j} - \mu_{1j}$ and $\mu_{3j} - \mu_{1j}$ by low order polynomials in j. Both analyses conclude that there are significant differences between the three response profiles, with $\mu_{1j} > \mu_{2j} > \mu_{3j}$ throughout the experiment.

One feature of the data that Diggle's model apparently fails to accommodate is that the *observed* mean response profiles tend to rise towards the end of the experiment. To allow for this, we extend model (29) to

$$\mu_{ij} = \begin{cases} \gamma_i - j\alpha & j < 3, \\ \gamma_i - 3\alpha + \eta(j-3) + \xi(j-3)^2 & j \geqslant 3. \end{cases} \tag{30}$$

For the covariance structure of the generic complete measurement sequence $\mathbf{Y}^*$ we assume equations (15) and (16) but with $\nu = 0$. Thus,

$$\operatorname{cov}\{Y_j^*, Y_k^*\} = \begin{cases} \sigma^2 + \tau^2 & j = k, \\ \sigma^2 \exp\{-\rho(j-k)^2\} & j \neq k. \end{cases}$$

Diggle (1990) used this model without the constraint $\nu = 0$ but estimated $\hat{\nu} \simeq 0$. It is at first sight unusual that the component of variation between animals is so small. A possible explanation is that inherent variability between cows is dominated by variation over time in factors common to all cows, e.g. weather or pasture quality. To complete our model we specify the drop-out probabilities as 0 up to and including week 14. We allow the underlying drop-out process to differ among weeks 15–19 but, because there are no drop-outs at week 18, the corresponding intercept parameter is extrinsically aliased and therefore the probability of drop-out on this occasion is set to 0. This leads to the logistic expression

$$\text{logit}\{p_k(H_k, y_k; \boldsymbol{\beta})\} = \beta_{0,k-14} + \beta_1 y_k + \beta_2 y_{k-1}, \qquad k = 15, 16, 17, 19. \quad (31)$$

Thus, β_1 and β_2 are the logistic regression coefficients for current and immediately previous protein contents respectively. One question of particular interest to us is whether the drop-out process is RD or ID, i.e. whether $\beta_1 = 0$ in equation (31). The maximum likelihood parameter estimates and minus twice the maximized log-likelihood from the CRD, RD and ID models are presented in Table 4. From these results we see that the likelihood ratio for the hypothesis $\beta_1 = 0$ is $2402.7 - 2388.5 = 14.2$ on 1 degree of freedom, which is very highly significant. We conclude that the drop-out process is ID.

Interestingly, and at first sight paradoxically, the maximum likelihood estimates of β_1 and β_2 have opposite sign, $\hat{\beta}_1 = 5.65$ and $\hat{\beta}_2 = -12.01$. Furthermore, although there is a strong negative correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$, the likelihood-based 95% confidence region for (β_1, β_2) is largely contained within the positive–negative quadrant, i.e. the region of the parameter space where $\beta_1 > 0$ and $\beta_2 < 0$. A reparameterization of the drop-out parameters leads to an easier interpretation of

TABLE 4
Milk protein trial: maximum likelihood estimates†

Parameter	*Maximum likelihood estimates for the following models:*				
	ID(int)	*ID(app)*	*RD*	*ID($\beta_2 = 0$)*	*CRD*
γ_1	4.15	4.15	4.15	4.15	4.15
γ_2	4.05	4.05	4.05	4.05	4.05
γ_3	3.93	3.93	3.93	3.93	3.93
α	−0.23	−0.23	−0.23	−0.23	−0.23
$\eta \times 100$	0.51	0.05	0.72	0.98	0.70
$\xi \times 100$	−0.02	−0.02	−0.05	−0.09	−0.05
β_{01}	18.98	19.64	19.36	13.21	−1.07
β_{02}	17.65	18.32	18.60	12.62	−1.71
β_{03}	16.59	17.20	17.80	12.08	−2.44
β_{04}	‡	‡	‡	‡	‡
β_{05}	18.28	18.89	18.54	12.11	−2.10
β_1	5.65	5.53		−4.46	
β_2	−12.01	−12.08	−6.29		
σ^2	0.066	0.066	0.070	0.075	0.069
τ^2	0.027	0.027	0.026	0.025	0.026
ρ	0.86	0.86	0.86	0.86	0.86
2 log-likelihood	2402.7	2402.7	2388.8	2351.5	2258.37

†ID(int), ID using numerical integration; ID(app), ID using the approximation to the integral.
‡Extrinsically aliased.

this result. Write equation (31) as

$$\text{logit}\{p_k(H_k, y_k; \boldsymbol{\beta})\} = \beta_{0,k-14} + \theta_1(y_k + y_{k-1}) + \theta_2(y_k - y_{k-1}), \qquad k = 15, 16, 17, 19. \tag{32}$$

Thus $\theta_1 = (\beta_1 + \beta_2)/2$ and $\theta_2 = (\beta_1 - \beta_2)/2$. The parameters θ_1 and θ_2 are more easily interpretable than are β_1 and β_2 since they represent dependence on level and increment in the response variable, and these quantities are likely to be much less strongly correlated than are y_k and y_{k-1}. The maximum likelihood estimates of the θ_j are $\hat{\theta}_1 = -3.18$ and $\hat{\theta}_2 = 8.83$, suggesting that the probability of drop-out increases when either the prevailing level of protein content is low or when the increment between the last and current protein contents is high.

Turning now to the interpretation of the mean response profiles μ_{ij}, the likelihood ratio statistic to test the hypothesis that $\eta = \xi = 0$ in model (30) is 0.61, which is not significant, i.e. we have no evidence in favour of a rise in μ_{ij} towards the end of the experiment. The reconciliation between this and the rise in the *observed* mean response lies in the drop-out process. We have seen that the probability of drop-out is higher among cows with a relatively low protein content. Furthermore, the sequence of protein contents for any one cow is strongly autocorrelated. The maximum likelihood estimate of μ_{ij} is, in effect, fitted from the conditional distribution of Y_j^* given earlier Y_k^* and our conclusion is that, for $j \geqslant 3$, *the conditional expectation of* Y_j^* *given its measurement history is a constant*. This is entirely consistent with a *rise* in the *conditional expectation of* Y_j *given non-drop-out*, which is what is being estimated by the *observed* mean response.

Finally, we note that our detailed consideration of the drop-out process for these data has not led to any material change in the conclusions about the complete measurement process $\{Y_j^*\}$. From a theoretical point of view, this is not surprising because the drop-outs are confined to the last five weeks of the experiment, and the drop-out-free phase of the experiment contains most of the information about the $\{Y_j^*\}$ process. From a practical point of view we think that it has been useful to expose the distinction between the mean response profiles for $\{Y_j^*\}$ and for Y_j conditional on non-drop-out.

5.2. *Mastitis in Dairy Cattle*

Mastitis is an infection of the udder that has serious economic consequences in the dairy industry. It reduces the milk yield of infected animals and persistent infection can lead to an animal's being removed prematurely from the herd. There is a view among dairy scientists, widely held, that mastitis is more likely to occur in high yielding cows. It may, for example, be the increased stress that such animals undergo that predisposes them to the infection. However, it is difficult to examine such a relationship because of the effects of mastitis. Although it has been established that there is a relationship between yield in one year and the occurrence of mastitis in the following year, it is of great interest to know how the occurrence of mastitis is related to the yield that *would have been obtained* in the same year as infection *had infection not occurred*. In the present context we can regard this as a missing value problem: the yield of an infected animal is treated as missing.

To investigate the relationship between yield and mastitis we use data on the total milk yields (in thousands of litres) for cows from a single herd in two consecutive

years. In each of five years, we select a group of cows who are in their third lactation and are free of mastitis. The data were obtained from the DAISY database in the Department of Agriculture at the University of Reading. To avoid complications with reinfection we include only animals with no previous history of infection. Of 107 animals, 27 (25%) became infected in their second year. There are many potential covariates that could be included in modelling such data. Here, to keep the analysis as simple as possible, we include only one: the year in which the animal was selected. From year to year the milk yield of the animals changes. Animals are therefore grouped according to the year of selection. The yield also tends to increase between the third and fourth lactations. We should therefore expect a change in average yield (assuming no mastitis) between the two years. We use the following linear model:

$$\mu_{ij} = \begin{cases} \gamma_i & j = 1, \\ \gamma_i + \delta & j = 2 \end{cases}$$

for an animal in year group i ($i = 1, \ldots, 5$). The parameters γ_i represent the means of the year groups in their first year, whereas δ represents the increment between the third and fourth lactations.

Since there are only two repeated measurements in this example we leave the covariance structure unconstrained. There are therefore three covariance parameters: the variances in years 1 and 2 (σ_t^2: $t = 1, 2$) and the correlation within the pairs of measurements (ρ).

We use the same logistic model for the drop-out probabilities as before, in this case noting that drop-out can only occur on one possible occasion, the second. Thus:

$$\text{logit}\{p_2(y_1, y_2; \boldsymbol{\beta})\} = \beta_0 + \beta_1 y_2 + \beta_2 y_1 .$$

In Table 5 we present the maximum likelihood estimates of the parameters from this model under the alternative drop-out models. Here the evidence for an ID process is only borderline. Comparing the log-likelihoods from the ID and RD models we obtain a likelihood ratio statistic of 4.94, which corresponds to the upper 3% point of the

TABLE 5
Mastitis in dairy cattle: maximum likelihood estimates†

Parameter	*Maximum likelihood estimates for the following models:*				
	ID(int)	*ID(app)*	*RD*	*ID($\beta_2 = 0$)*	*CRD*
γ_1	5.80	5.80	5.82	5.83	5.83
γ_2	5.59	5.59	5.61	5.61	5.61
γ_3	5.96	5.94	5.94	5.94	5.94
γ_4	5.88	5.87	5.92	5.93	5.92
γ_5	5.63	5.61	5.58	5.57	5.58
δ	0.32	0.32	0.72	0.81	0.72
β_0	0.15	0.13	−2.64	−3.81	−1.09
β_1	−2.63	−2.60		0.41	
β_2	2.38	2.33	0.27		
σ_1^2	0.83	0.83	0.83	0.83	0.83
σ_2^2	1.67	1.66	1.33	1.37	1.33
ρ	0.48	0.48	0.59	0.59	0.59
2 log-likelihood	−270.94	−271.05	−275.88	−275.72	−277.11

†ID(int), ID using numerical integration; ID(app), ID using the approximation to the integral.

χ^2_1-distribution, and comparing the RD and CRD models we obtain a statistic of only 1.23.

If we accept that an ID process is operating we again need to consider the form that it is taking. As with the protein data, $\hat{\beta}_1$ and $\hat{\beta}_2$ have opposite signs although in this case it is $\hat{\beta}_1$ that has the negative sign. Again, a likelihood-based 95% confidence region for these two parameters lies largely in the same quadrant. We can reparameterize the drop-out probabilities as in equation (32) to obtain

$$\begin{aligned}\text{logit}\{p_k(y_1, y_2; \boldsymbol{\beta})\} &= \beta_0 + \theta_1(y_1 + y_2) + \theta_2(y_2 - y_1) \\ &= 0.15 - 0.125(y_1 + y_2) - 2.505(y_2 - y_1).\end{aligned}$$

We see that the probability of drop-out is increased with a *decrease* in yield from the first to second years, whereas the dependence on the average yield of the animal over the two years is comparatively small. From this, we would expect that the RD model ($\beta_1 = 0$) and ID model with $\beta_2 = 0$ would fit little better than the CRD model in which both parameters are omitted, and the results in Table 4 bear this out. Only with both parameters in the model can we model dependence on change from one year to the next. The drop-out parameter estimates from the milk protein example (Table 4) behave similarly, except that there the dependence on average response is stronger and so both models with dependence on a single response provide a markedly better fit than the CRD model.

We consider now the estimates of the linear model parameters. The estimates of the γ_i are approximately constant across all the drop-out models. Since these measure the mean yield in the first year they should be largely independent of the model for the drop-out mechanism. However, the increment between years, measured by δ, is noticeably smaller (0.32) under the ID model than under the RD and CRD models (0.72). The difference of 0.40 between the estimates represents 400 l, a value of practical significance. The difference can be explained in terms of the parameters of the drop-out component of the model. The estimated difference between β_1 and β_2 implies that animals with a marked decrease from the first to the second year are more likely to contract mastitis and so the uncorrected (conditional) means from the non-mastitic cows in the second year will overestimate the underlying marginal mean, hence the larger $\hat{\delta}$ under the RD and CRD models. Note also the smaller variance of the non-mastitic yields in the second year (1.33) compared with the estimated marginal variance (1.67). This is another manifestation of the selection process represented by the ID mechanism which is not taken into account in the CRD and RD analyses.

In terms of the relationship between the occurrence of mastitis and milk yield these results suggest that it is not necessarily just the occurrence of high yield that is predisposing an animal to mastitis but previous high yield relative to that animal's typical or expected yield. We accept that, in the light of the borderline nature of the results and the rather crude models involved, these can be at best tentative conclusions. However, the results warrant further investigation, and these analyses are being pursued with larger sets of data and with somewhat more sophisticated models.

5.3. *Antidepressant Trial*

Our third example is taken from a multicentre clinical trial on the treatment of depression. In each of six centres subjects were randomized to one of three treatments,

TABLE 6
Numbers of drop-outs in the antidepressant trial

Week	*Centre*																	
	1			*2*			*3*			*4*			*5*			*6*		
	Treatment																	
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	3	0	3	4	3	3	6	0	1	4	3	1	1	3	4	8	3
3	2	3	1	4	1	2	1	1	0	2	2	2	0	2	2	0	0	2
4	4	2	7	3	0	0	1	1	2	2	1	3	4	3	0	7	0	3
Total	8	8	8	10	5	5	5	8	2	5	7	7	5	6	5	11	8	8
Total	24			20			15			19			16			27		

approximately 20 subjects receiving each treatment in each centre. 369 subjects entered the trial. Each subject was rated on the Hamilton depression score HAMD, a sum of 16 test items producing a response on a 0–50 scale. Measurements were made on each of five weekly visits, the first made before treatment, the remaining four during treatment. We refer to these as weeks 0–4. Subjects dropped out of the trial from week 2 onwards and, by the end, 121 (33%) had left. In Table 6 we present the numbers of drop-outs by week, centre and treatment. It can be seen that drop-outs occur for all treatments and centres but on three occasions a substantial proportion (over a third) of the subjects leave a particular treatment group in a particular centre.

A subset of these data has been analysed by Heyting *et al.* (1990), who considered several analyses, including a maximum likelihood analysis under an antedependence covariance structure. In the following we adopt the covariance structure of Heyting *et al.* (1990) but use a less general linear model. The assumption of an antedependence covariance structure of order r (AD(r)) is equivalent to the requirement that Y_j^* is conditionally independent of $Y_{j-r-1}^*, \ldots, Y_1^*$ given $Y_{j-1}^*, \ldots, Y_{j-r}^*$, which, in turn, is equivalent to the requirement that the inverse of the covariance matrix is banded with r non-zero diagonals above and below the main diagonal. With equally spaced measurements such a structure can be derived from a general non-stationary autoregressive process of order r, and the familiar stationary AR(r) structure is a special case of the AD(r) model. Heyting *et al.* (1990) establish that an AD(2) structure is appropriate for these data and we adopt this in the following.

We now consider a model for the mean profiles. In Fig. 2 we present the set of mean profiles for each combination of treatment and centre. As in the milk protein example these are the simple average profiles based on the observed data. Unless we have a CRD process these may be biased. It is clear from the plots that there is an overall decrease in HAMD score and it is this trend that we wish to model for each treatment, and to compare between treatments. We know from the initial randomization that, within each centre, the treatment profiles will have common intercepts at week 0. We therefore use a model in which each centre is allowed to have a different intercept and, to accommodate possible non-linearity in the time profiles, we use quadratic regression relationships for each treatment group. Thus we have the model

$$\mu_{cij} = \gamma_c + \eta_i j + \xi_i j^2 \qquad j = 0, \ldots, 4 \tag{33}$$

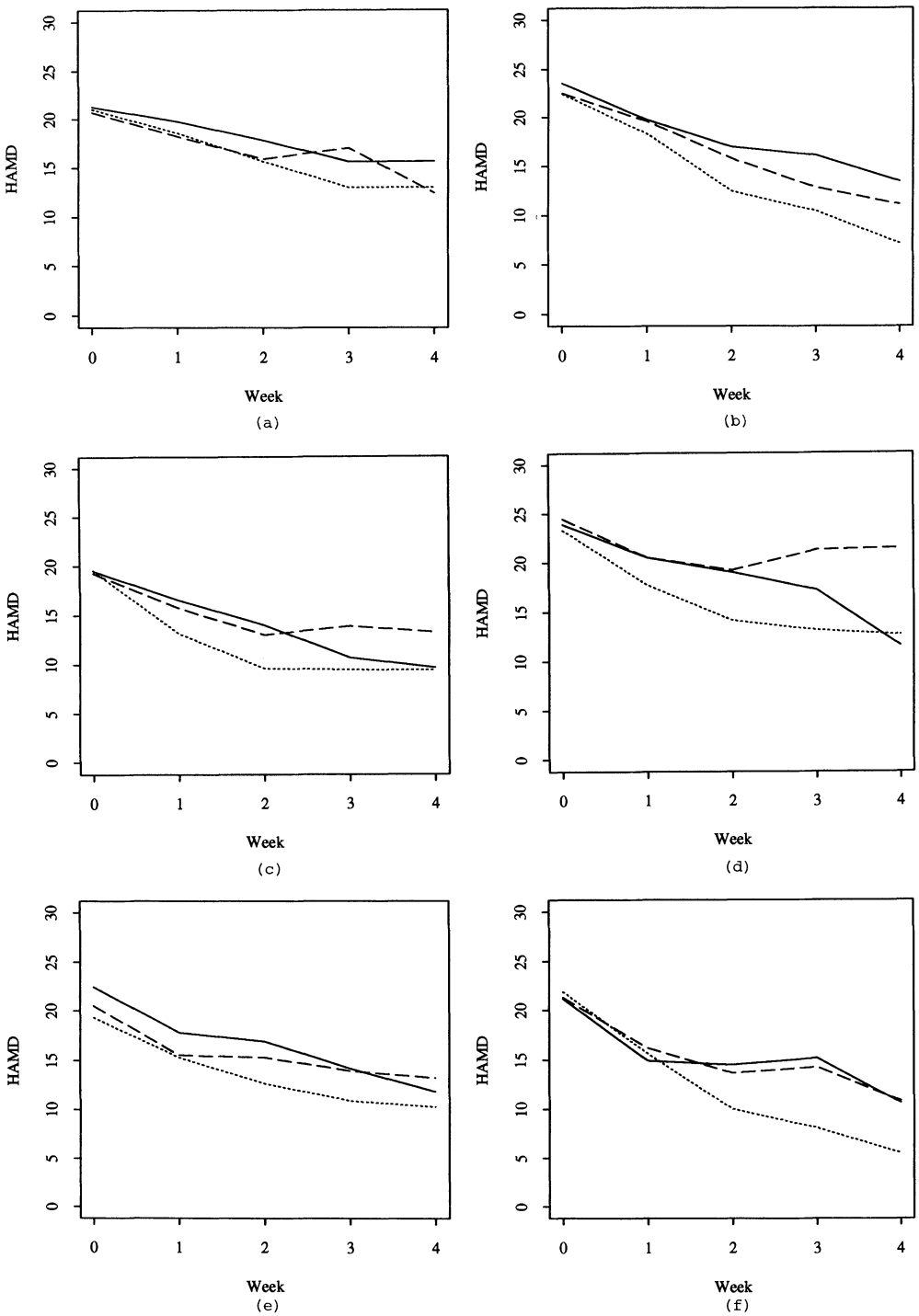

Fig. 2. Observed mean response profiles from the antidepressant trial (———, treatment 1; – – –, treatment 2; · · · · · ·, treatment 3): (a) centre 1; (b) centre 2; (c) centre 3; (d) centre 4; (e) centre 5; (f) centre 6

for the expectation of the observation from a subject on the jth visit in treatment group i and centre c. Note that we have not allowed for interaction between treatment effect and centre in this model. Although Heyting *et al.* (1990) find some evidence for such an interaction, we continue here with the simpler model. The addition of the interaction term increases considerably the number of parameters in the model and the behaviour of the simplex algorithm used to maximize the likelihood deteriorates rapidly with increasing dimensionality of the problem. The addition of the interaction term would not, we conjecture, change substantially the conclusions that emerge from our analysis.

We use the same logistic regression model (32) for the occurrence of drop-out as in the previous two examples. In this case the underlying drop-out rate is set to 0 for week 1, when there are no drop-outs, and constant for weeks 2–4. Allowing this parameter to vary over time produces only a negligible change in the value of the maximized likelihood.

The maximum likelihood estimates for the linear model and drop-out parameters are presented in Table 7. Comparing the maximized log-likelihoods from the ID, RD and CRD models we see that there is strong evidence for ID; twice the log-likelihood difference between the RD and ID models is 8.3 on 1 degree of freedom. As with both previous examples the parameters β_1 and β_2 have opposite signs; in this case β_2 is positive. This points to a strong association between drop-out and the increment in HAMD score between two successive visits.

The choice of drop-out model has some effect on the estimates of the linear model parameters. As expected, the estimates of the centre parameters $\hat{\gamma}_i$, corresponding to intercepts at week 0, are comparatively stable with respect to the drop-out model. Although the effect on the linear and quadratic parameter estimates of moving from the RD to ID model is not consistent among treatments, the combined effect on the

TABLE 7
Antidepressant trial: maximum likelihood estimates†

Parameter	*Maximum likelihood estimates for the following models:*				
	ID(int)	*ID(app)*	*RD*	*ID($\beta_2 = 0$)*	*CRD*
γ_1	21.38	21.38	21.38	21.35	21.38
γ_2	22.53	22.51	22.57	22.61	22.56
γ_3	19.48	19.47	19.41	19.41	19.43
γ_4	23.99	24.00	24.00	24.98	23.97
γ_5	20.94	20.92	20.86	20.84	20.88
γ_6	21.11	21.07	21.20	21.27	21.20
η_1	−3.42	−3.38	−3.42	−3.41	−3.35
η_2	−5.78	−5.73	−5.27	−5.12	−5.33
η_3	−3.83	−3.83	−3.70	−3.62	−3.67
ξ_1	0.20	0.20	0.30	0.35	0.29
ξ_2	0.67	0.67	0.64	0.64	0.65
ξ_3	0.47	0.47	0.52	0.53	0.51
β_0	−3.30	−3.50	−3.29	−3.27	−1.95
β_1	−0.23	−0.22		0.08	
β_2	0.25	0.25	0.08		
2 log-likelihood	−7047.0	−7047.5	−7055.3	−7063.3	−7085.1

†ID(int), ID using numerical integration; ID(app), ID using the approximation to the integral.

profiles is similar for each treatment: the marginal (ID) model profiles decrease more rapidly than the conditional profiles (RD and CRD models). This we should expect, given the signs and sizes of the estimated parameters in the various drop-out models. Subjects with greater negative increments in HAMD score between visits are more likely to drop out.

It is interesting that although there is clear evidence of differences between the treatments under the ID model (the likelihood ratio statistic is 32.4 on 4 degrees of freedom for the null hypothesis of no treatment differences between the profiles) the treatment differences between the profiles are quite consistent across drop-out models. As a description of the *comparisons between* treatment groups the estimates from the RD and CRD models serve reasonably well for both the marginal process $\mathbf{Y}^*$ and the conditional process $\mathbf{Y}$, but as a description of the *mean marginal profile* for a particular treatment group the profiles produced by the RD and CRD models are misleading. However, it might be argued that, given the model for the drop-out process, we should not expect to see large differences between treatment effects among the different drop-out models: the treatments can only influence the drop-out rate through their effect on the response. The relationship could be made more direct by, for example, introducing treatment effects directly into the model for the drop-out process, either as constants or by allowing the relationship between response and drop-out, as currently defined by β_1 and β_2, to differ between treatments. Such generalizations are included in the general class of model developed in this paper and we return to these issues in the discussion.

6. Discussion

In this paper we have developed a modelling framework for ID and as a first step we have concentrated on models in which there is a simple relationship between drop-out probability and response. Within the framework developed, it is straightforward in principle to generalize this relationship and there are many circumstances in which one might wish to consider such generalizations. Drop-out probabilities may be allowed to depend on unit-specific or time-dependent covariates in a direct way, as for example we allow dependence on time in the milk protein example. One might also consider such relationships in the antidepressant trial in which the occurrence of side-effects is likely to affect drop-out and measurements of these are available at each visit. It is quite possible that the inclusion of such covariates may change an ID to an RD process. If such covariates affect both the drop-out probability and response then, in the absence of the covariate, the response itself may show a relationship with drop-out. One might consider including covariates with the purpose of reducing or eliminating dependence of drop-out probabilities on missing observations. Further generalizations might allow interactions between covariates and dependence of drop-out probability on response.

We remark that all such generalizations, although straightforward in principle, would require sufficient information to be available in the data. We may find in the more complicated models or in the smaller data sets that terms in the drop-out component of the model are effectively aliased. An inspection of the likelihood surface would shed light on this. For such modelling we would also need to use a more sophisticated numerical procedure than the simplex algorithm. Although this proved sufficiently robust for the problems tackled, with the larger models (those involving 20 or more

parameters) convergence was painfully slow. Computation of the likelihood itself was considerably faster by using the approximation described in Section 3, and the resulting parameter estimates were very similar to those obtained from the exact likelihood. This is to be expected, given that it is generally difficult to distinguish between logit and probit models except with extremely large data sets. However, the approximation does not avoid the need for a general optimization algorithm. It is likely that a procedure such as the Newton–Raphson method with numerical second derivatives will be necessary to use this modelling approach in its full generality. The form of likelihood itself does not suggest any obvious analytical simplifications or short-cuts that could be applied to the numerical maximization.

The simplex algorithm does not provide an estimate of the information matrix for the parameter estimates. If inferences were to be based on the information matrix an estimate could be obtained from numerical second derivatives of the maximized likelihood. However, our experience has been that exploration of the likelihood surface is an important component of such modelling, and this leads more naturally to the use of inferences based on likelihood ratio tests and likelihood-based confidence regions. Recent developments in Monte Carlo inference have opened up new possibilities for exploring likelihood surfaces for complex models. See, for example, Geyer and Thompson (1992), Smith and Roberts (1993), Besag and Green (1993), Gilks *et al.* (1993) and the associated discussion of these papers.

The simulation study described in Section 4 has shown that the proposed method works as expected, eliminating the bias that appears in OLS and RD likelihood analyses that ignore ID mechanisms. These results also suggest that the RD analysis provides some protection against the ID bias compared with the OLS analysis. This protection increases with increasing within-subject correlation and can be explained by previous observations acting as surrogates for the missing observation. The greater the mutual dependence the greater the effectiveness of the surrogate.

We have deliberately avoided the formulation of a specific modelling strategy. The approach described above could be incorporated into such strategies in a variety of ways and any particular choice must depend on the context. The relationship between drop-out and missing values might be of interest in its own right, such as in the mastitis example. Alternatively such models might be considered in the context of an investigation of sensitivity of inference to the drop-out process.

The method proposed relies on the correct description of the conditional distribution of a missing value given the history of the process. This suggests that the estimators may not be robust to misspecification of the joint distribution, particularly the covariance structure. Interestingly, earlier analyses of the data from the antidepressant trial using a uniform covariance structure still pointed strongly to an ID process even though, in this example, the uniform structure is clearly incompatible with the data.

All three examples provided evidence of an ID process and in two of these cases the evidence was strong. All three also produced parameter estimates for the drop-out process that took opposite signs for the current and previous observations, in each case pointing to a relationship between the incremental change and the probability of drop-out. This suggests that ID may be the rule rather than the exception.

One major unresolved issue is the question of how to deal with ID in *categorical* longitudinal data. Stasny (1990) and Conaway (1993) develop a model based on the Markov chain that incorporates ID. One alternative approach would be to augment the set of possible values of the response by an absorbing state which represents

drop-out. Another would be to use parameter-driven models (Zeger, 1988). For example, for a binary response, let $p(t)$ denote the probability of a positive response at time t. Then, we might postulate that the responses on a given experimental unit are conditionally independent given the corresponding $p(t)$, but that $Y(t) = \log[p(t)/\{1-p(t)\}]$ is a stochastic process of the kind considered in this paper. This kind of model is closely related to non-Gaussian Kalman filtering (Kitagawa, 1987), with the model for $Y(t)$ corresponding to the state equation and the conditionally independent response sequence to the observation equation. Whatever modelling approach is adopted, we forsee a need to overcome two major obstacles. First, likelihood-based methods for correlated categorical data usually raise considerable technical difficulties. Second, and more fundamentally, the coarser scale of measurement embodied in categorical, as opposed to continuous, response variables limits the information that the data can convey about any postulated relationship between drop-out and measurement history.

Acknowledgements

We are grateful to Hans Essers and the Pharmaceutical Division of Solvay Duphar BV for permission to use the data from the psychiatric study and to Dr R. Esselmont and the Farm Management Department of the University of Reading for permission to use the data on mastitis in dairy cattle. We also thank the referees for their comments on an earlier version of this paper.

References

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc.* B, **55**, 25–37.

Chatfield, C. and Collins, A. J. (1980) *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Conaway, M. R. (1993) Non-ignorable non-response models for time-ordered categorical variables. *Appl. Statist.*, **42**, 105–116.

Cullis, B. R. and McGilchrist, C. A. (1990) A flexible model for the analysis of growth data from designed field experiments. *Biometrics*, **46**, 131–142.

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

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

——(1990) *Time Series: a Biostatistical Introduction*. Oxford: Oxford University Press.

Gabriel, K. R. (1961) The model of ante-dependence for data of biological growth. *Bull. Inst. Int. Statist.*, **39**, 253–264.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J. R. Statist. Soc.* B, **55**, 39–52.

Glynn, J. J., Laird, N. M. and Rubin, D. B. (1986) Selection modelling *versus* mixture modelling with nonignorable nonresponse. In *Drawing Inferences from Self Selected Samples* (ed. H. Wainer), pp. 115–142. New York: Springer.

Gould, A. L. (1980) A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics*, **36**, 721–727.

Greenlees, W. S., Reece, J. S. and Zieschang, K. D. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *J. Am. Statist. Ass.*, **77**, 251–261.

Heckman, J. J. and Singer, B. (1985) *Longitudinal Analysis of Labour Market Data*. Cambridge: Cambridge University Press.

Heyting, A., Essers, J. G. A. and Tolboom, J. T. B. M. (1990) A practical application of the Patel–Kenward analysis of covariance to data from an anti-depressant trial with drop-outs. *Statist. Appl.*, **2**, 295–307.
Jones, R. H. and Ackerson, L. M. (1990) Serial correlation in unequally spaced longitudinal data. *Biometrika*, **77**, 721–731.
Jones, R. H. and Boadi-Boteng, F. (1991) Unequally spaced data with serial correlation. *Biometrics*, **47**, 161–175.
Kenward, M. G. (1987) A method for comparing profiles of repeated measurements. *Appl. Statist.*, **36**, 296–308.
Kitagawa, G. (1987) Non-Gaussian state-space modelling of nonstationary time series. *J. Am. Statist. Ass.*, **82**, 1032–1063.
Laird, N. M. (1988) Missing data in longitudinal studies. *Statist. Med.*, **7**, 305–315.
Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
Munoz, A., Carey, V., Schouten, J. P., Segal, M. and Rosner, B. (1992) A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics*, **48**, 733–742.
Murray, G. D. and Findlay, J. G. (1988) Correcting for bias caused by drop-outs in hypertension trials. *Statist. Med.*, **7**, 941–946.
Nelder, J. A. and Mead, R. (1965) A simplex method for function minimisation. *Comput. J.*, **7**, 303–313.
Pantula, S. G. and Pollock, R. A. (1985) Nested analysis of variance with autocorrelated errors. *Biometrics*, **41**, 909–920.
Pledger, G. and Hall, D. (1982) Correspondence: withdrawal from drug trials. *Biometrics*, **38**, 276–278.
Ridout, M. S. (1991) Reader reaction: testing for random dropouts in repeated measurements data. *Biometrics*, **47**, 1255–1258.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Shih, W. J. (1992) On informative and random dropouts in longitudinal studies. *Biometrics*, **48**, 970–972.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Stasny, E. (1990) Symmetry in flows among reported victimization classifications with nonrandom nonresponse. *Surv. Methodol.*, **16**, 305–330.
Verbyla, A. P. and Cullis, B. R. (1990) Modelling in repeated measures experiments. *Appl. Statist.*, **39**, 341–356.
Wang, F., Laird, N. M. and Ware, J. H. (1992) A simulation study of estimators for rates of change in longitudinal studies with attrition. To be published.
Ware, J. H. (1985) Linear models for the analysis of longitudinal studies. *Am. Statistn*, **39**, 95–101.
Williams, E. J. (1949) Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. Sci. Res.* A, **2**, 149–168.
Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.
Zeger, S. L. (1988) A regression model for time series of counts. *Biometrika*, **75**, 621–629.
Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060; correction, **45** (1989), 347.

Discussion of the Paper by Diggle and Kenward

J. N. S. Matthews (University of Newcastle upon Tyne): You will observe that the paper's running title is 'Diggle and Kenward Informative Drop-out'. I shall refrain from further comment on the implied identifiability problem, except to say that both authors have been exceedingly informative.

The analysis of longitudinal data is one of the most common to confront the statistician, at least in the biomedical fields, and missing data must be a feature of most of these data sets. Despite very many papers on longitudinal data analysis, the relative paucity of papers dealing with missing data suggests that methodologists are reluctant to tackle the problem. Indeed, too many papers and text-books state that a method of longitudinal data analysis 'accommodates missing data' when they really mean that it accommodates unbalanced data.

Missing data are an embarrassment to both data collector and analyst. Missing data often make the data collector feel incompetent and guilty, a reaction I confess to doing little to disabuse as it is a useful cover for my own unease. Although all practising statisticians realize the importance of ascertaining *why* data are missing, the fact remains that we seldom know what to do next. The completely random drop-out (CRD), random drop-out (RD), informative drop-out (ID) hierarchy proposed by Rubin was always conceptually useful but lacked a convincing implementation for the case that really mattered, namely ID. We should be grateful to Diggle and Kenward as they have provided such an implementation.

With a method which will find application in many fields, the authors could not describe more than a few of the ways that the model could be used. There is much potential to unlock in equation (20), where the drop-out process is modelled with time-dependent variables other than the main response. An example arises in clinical trials (most of which collect longitudinal data) where, even though the response to treatment may be adequate (or better), withdrawal is often determined by the severity of side-effects, which are usually measured by variables other than the primary response.

While on the subject of clinical trials, it is appropriate to mention the dictum of 'analysis by intention to treat', a device which has done more than most to bewilder clinical colleagues. Such analyses are essential if the benefits of randomization are to be preserved in the face of an unco-operative world. Nevertheless, the results obtained are second best—the responses of patients who can tolerate treatment being mixed with those who cannot. In avoiding bias we avoid some of the most relevant clinical questions. I would be interested in the authors' views on the contribution that their method can make, not only to analysis but also to the timing of observations. Here, and elsewhere, it will be of great importance to decide on whether inference should be based on $\mathbf{Y}$ or $\mathbf{Y}^*$: it should be considered an advantage that the method requires great clarity about which questions are to be answered.

On a general level, the method faces a seemingly inescapable problem. If there are many drop-outs, with a substantial proportion at an early stage, the proposed methods *can* be applied, but how many of us would feel happy to rely on technical virtuosity when, say, 60% of the data are absent? This may reflect my lack of virtue but I suspect that there would be more than a little sympathy for the attitude in a profession as notoriously cautious as ours. Alternatively, if the proportion of drop-outs is low, then much less can be learnt about the drop-out process; this gives low power to discriminate between drop-out processes, and in the ID case large variances for $\hat{\beta}$ may inflate the variances of $(\hat{\theta}, \hat{\phi})$.

A partial resolution of this is suggested by the intriguingly good performances reported in Tables 1–3 of RD analyses of ID processes, even for drop-out rates up to 50%. Although the authors' preference for likelihood-based confidence intervals is appropriate for the examples, a proper assessment of the simulation studies was frustrated as the authors did not report mean-square or standard errors. For low to moderate drop-out rates, estimating parameters of the drop-out process may mean that the ID analysis has a larger mean-square error and the RD analysis may be preferable. As an aside, the appendix to Nelder and Mead's (1965) paper shows how the simplex method can provide estimates of the observed information matrix (but see also Brumby (1989) and Phillips and Eyring (1988)); O'Neill's (1971) algorithm does not provide this facility but an alternative which does (by Shaw, with Wedderburn and Miller) is held on STATLIB.

Extensions of the method, e.g. to intermittent missing values and methods for testing the underlying assumptions (which I suspect are rather strong), are obviously needed. So too is much wider practical and simulation experience but the first steps are always the most difficult. The authors must be heartily congratulated for providing an invaluable approach to a common and difficult problem, and for opening many areas that are ripe for further exploration and it is a great pleasure to propose our vote of thanks.

A. P. Grieve (ZENECA Pharmaceuticals, Macclesfield): The authors have presented a framework for modelling informative drop-outs in longitudinal data analysis and I welcome the paper as one more example of the use of more realistic models in the biological sciences. As with any new modelling framework several standard questions come to mind: is the framework adequate?; are there alternative models?; is the inferential apparatus satisfactory?

The adequacy of any modelling framework can only fully be tested in practice. The examples given are from agricultural and clinical research but there are other potential application areas such as chronic regulatory toxicology studies. In any application there will be many potential causes of drop-out. For example Heyting *et al.* (1992) give six common reasons for patients' withdrawal from a clinical study:

(a) recovery;
(b) lack of improvement;

(c) treatment-related side-effects;
(d) unpleasant study procedures;
(e) intercurrent health problems;
(f) external factors unrelated to the trial.

Either of the first two clearly rule out completely random drop-out (CRD). But it is not clear whether either cause is more likely to give RD rather than informative drop-out (ID) or vice versa. Are there any reasons for believing that a particular cause is more likely to give rise to one than another? Of course, in practice, there will be more than one cause of drop-out in a trial and this may affect the estimation of parameters of the drop-out process.

Two of the authors' three examples clearly point to ID rather than RD suggesting strongly that we need to determine the appropriate drop-out process. Therefore it was rather surprising that the simulation results showed that 'the biases seen in the CRD analyses, although consistent, were small and would not be of great practical concern'. Is it possible, as the authors allude to in Section 6, that inclusion of the appropriate covariates in the examples 'might change an ID to RD'? If so the question arises whether this generalization of a method originally proposed for use with non-response problems in surveys really is appropriate in longitudinal studies. In surveys the unobservable data exist; in the longitudinal case do they exist?

Accepting that the model is appropriate a couple of questions concerning the model for the drop-out process arise. The authors assume that in the most general model the conditional probability of drop-out at time d depends on the history of the measurement process up to and including time d, including the value which would be observed if the unit did not drop out. They remark, '"drop-out at t_d" actually means "drop-out somewhere between t_{d-1} and t_d"'. If a unit drops out soon after t_{d-1} and $t_d - t_{d-1}$ is long then it seems to me that the conditional probability should depend on the value of the measurement process at the time of drop-out rather than at the nominal measurement time. To what extent will there be bias if this conjecture is true? Secondly, the authors provide a motivation for their logistic formulation by observing that it may be reasonable to assume that the probability of drop-out depends on a discounted integral of the history of the measurement process with the expression $\beta_1 y + \Sigma_{j=2}^{k} \beta_k y_{k+1-j}$ being interpretable as a quadrature approximation to the integral. It would interest me to know to what extent this approximation is robust to the width of the observation interval.

The authors comment that, depending on the application, any one of the mean structures of $\mathbf{Y}$, $\mathbf{Y}^*$ or $\mathbf{Y}$ given no drop-out may be of interest. To this list I would add the mean structure of $\mathbf{Y}^*$ given drop-out. In Fig. 3 are displayed the barley diet data taken from the milk protein trial in which the data from drop-outs are distinguished from the complete profiles. It is apparent from this figure that already at week 11 there is an indication that the drop-outs have lower protein content values than the

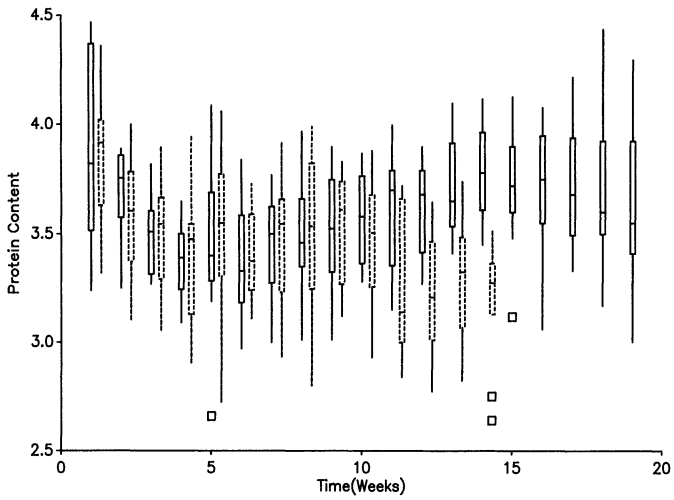

Fig. 3. Conditional response profiles of protein content in the barley diet group: ———, completers; ---------, drop-outs

completers. That information may be as important for the agronomist to know as it would be to a toxicologist if the data were animal weights in a chronic toxicology study.

In the terminology of Rubin (1987) the method of this paper falls in the class of 'selection models'. Rubin considers a second model to handle what he terms 'non-ignorable non-response' which he calls a 'mixture model'. As its name implies, this model involves separate distributions for responders and non-responders. I doubt whether this approach can be generalized to the longitudinal case as it will be necessary to specify a separate distribution for each drop-out time point and this will rapidly lead to an explosion in the number of mean value parameters. I would be interested to know whether the authors have looked at other methods of modelling informative drop-out.

Turning to the inferential apparatus I was disappointed that the authors hardly went beyond point estimation and likelihood ratio tests for model comparison. Minimally it would be wise to determine the asymptotic standard errors of the estimated parameters as they suggest either numerically following a simplex maximization or as part of the maximization process itself. If there is a substantial difference between the likelihood ratio test for a parameter β and the asymptotic test based on $\{\hat{\beta}/\mathrm{se}(\hat{\beta})\}^2$ then there is an indication that the likelihood surface is skewed in the β-direction. But surely modern likelihood theory has more to do with providing profiles of likelihoods rather than simply determining point estimates with associated errors. In a sense the authors accept this view by referring to the recent Markov chain Monte Carlo (MCMC) work and by talking of the use of confidence regions based on likelihood ratios. But in the paper even when they calculate a confidence region we are not shown it, only told that it lies in a particular quadrant of the plane thereby losing the information which its shape contains about for example the correlation between the parameter estimates or whether the likelihood surface is elliptical. In their oral presentation such a likelihood-based 95% region was shown and it clearly showed strong correlation and a very long nose in the direction of the origin.

The authors comment on the deterioration of the behaviour of the simplex algorithm with increasing dimensionality. Maximum likelihood is itself suspect under such circumstances as Diggle (1990) points out as it leads to biased covariance parameter estimates. Perhaps the authors could tell us whether they have considered the use of restricted maximum likelihood estimation techniques for their model which can considerably reduce the bias in the estimates of the covariance parameters. Or is this where the true advantage of MCMC methods lies?

I have enjoyed reading this paper and have great pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

C. J. Skinner (University of Southampton): This paper is related to work on non-response in surveys; for example the informative drop-out (ID) model resembles the model of Hausman and Wise (1979) (see also Hsiao (1986)). The need for caution in the use of such models because of non-robustness to distributional misspecification has, however, been demonstrated by Little and Rubin (1987), chapter 11. In particular, it seems difficult to disentangle informative non-response in a Gaussian model from simple distributional skewness. To consider this with the methods proposed I looked at one example.

Sets of data were simulated from the model $Y_{ij}^* = u_i + v_{ij}$, $u_i \sim N(\mu, \sigma_u^2)$, $v_{ij} \sim N(0, \sigma_v^2)$, $i = 1, \ldots, 100$, $j = 1, 2$, $\mathrm{logit}\{\Pr(D = 2|\mathrm{history})\} = \beta_0 + \beta_2 y_1 + \beta_1 y_2$. Misspecification was represented by truncating the distribution of u_i below at $\mu - \sigma_u$. The results revealed some misspecification effects but more striking was the high sampling variability of the β-estimates when assuming the ID model, even if the Gaussian model held. This variability translated in a less pronounced way into inflated standard errors for the estimated linear model parameters (see the comments of Matthews and Grieve).

For example, 1000 sets of data were generated from the Gaussian ID model with $\mu = 10$, $\sigma_u = 5$, $\sigma_v = 1$, $\beta_0 = -2$ and $\beta_1 = \beta_2 = 0.1$. Considering estimators of $E(Y_{i2}^* - Y_{i1}^*)$, the estimated bias of the estimator based on the ID model was 0.04, compared with the value -0.11 for the estimator based on the RD model. This accords with results in Tables 1–3. However, the standard deviation of the estimator based on the ID model was 0.80 compared with the value 0.22 for the random drop-out (RD) model, implying a mean-squared error (MSE) 10 times larger. If in addition there was distributional misspecification the MSE was 13 times larger. For a second example with $\mu = 10$, $\sigma_u = 1$, $\sigma_v = 1$, $\beta_0 = -10$ and $\beta_1 = \beta_2 = 0.5$, the MSE of the estimator based on the ID model was still three times larger than that for the estimator based on the RD model even when the ID model was correct.

This is very limited evidence, but what indication there is here of inflated standard errors and sensitivity to distributional assumptions together with reasoning, as in section 11.4 of Little and Rubin (1987), lead me to suggest that the longitudinal data which remain after drop-out may by themselves contain

very little information about the 'informativeness' of the drop-out. It is thus desirable if possible that external information about the drop-out mechanism be sought and used. If one is concerned that estimates based on the RD model may be unsatisfactory then it may be more sensible to make inference under alternative plausible *postulated* types and degrees of informativeness, as in a sensitivity analysis (see Little and Rubin (1987), section 11.5), rather than to extend the parametric model as proposed here.

R. Henderson (University of Newcastle upon Tyne): Diggle and Kenward are to be congratulated on their realistic approach to a genuine practical problem. I have three remarks.

The first concerns the drop-out process. Logistic models for grouped survival data are sometimes used in practice but it is perhaps more natural to assume that drop-out results from a continuous underlying proportional hazards model. Thus $Y_{ii}, \ldots, Y_{in}$ can be assumed to be noisy observations at $t_{ii}, \ldots, t_{in}$ on a continuous process $Y_i(t)$, perhaps a marker process acting as proxy for the 'health' of the subject (Jewell and Nielson, 1993). For history H assume that the hazard is given by $\lambda(t; H) = \lambda_0(t) \exp\{\theta\, Y(t)\}$ depending only on the current value of $Y(t)$, not the sample path. Then conditional on survival to t_{k-1} the probability of drop-out by t_k is

$$1 - \exp\left[- \int_{t_{k-1}}^{t_k} \lambda(u) \exp\{\theta\, Y(u)\}\, du \right].$$

Following Section 3.2 we might approximate this by the informative drop-out model

$$p_k(H_k, y; \beta) = 1 - \exp\{-g(y, y_{k-1}, \beta)\}$$

for some function g, as an alternative to model (18). In this formulation observations before t_{k-1} do not affect drop-out, as assumed in the examples in Section 5.

The second remark is that it would be useful to add profiles of expected mean $E[\, Y_k \hat{S}(t_k | \mathbf{Y}^{(k)})\,] / E[\, \hat{S}(t_k \mathbf{Y}^{(k)})\,]$ to plots like Fig. 1, where expectation is over the multivariate normal distribution of $\mathbf{Y}^{(k)} = (Y_1, \ldots, Y_k)^{\mathrm{T}}$ and $\hat{S}$ is the estimated survivor function obtained from model (18) or proportional hazards as previously mentioned. Numerical integration may be required but the plot could give an indication of model suitability, especially if the data could be split into training and validation samples. Is this feasible? Note that $\hat{S}$ can be obtained in a more simple form from the proportional hazards model than from the logistic.

The final point concerns influence. Long-term survivors despite poor prognosis often have particularly high influence in survival analysis (Henderson and Oman, 1993) and simulation results indicate that similar properties hold for the Diggle and Kenward model. I considered the two-period situation with $\mu_1 = \mu_2 = 0$, $\sigma^2 = \sigma_2^2 = 1$ and $\rho = 0.5$ all assumed to be known. Newton–Raphson iteration was used to estimate β_0, β_1 and β_2 from an original sample of $m = 100$ simulated observations and then the same sample augmented by a single specified added case. True values were $\beta_0 = -1$ and $\beta_1 = \beta_2 = 1$, and Table 8 gives mean estimates from batches of 500 simulations: note the serious underestimation of β_1 at added case $Y_1 = Y_2 = 3$ with no drop-out. This is something to watch out for.

G. E. P. Box (University of Wisconsin, Madison): It is an important step forwards to consider analysis which makes use of incidental information such as whether a subject drops out of the study. It is important to remember, however, that building appropriate models cannot be a purely mathematical exercise; it must be conducted in close co-operation with a subject-matter specialist. A conversation should be

TABLE 8
Mean coefficients from 500 simulations before (i) and after (ii) augmentation

New case		β_0		β_1		β_2	
y_1	y_2	*(i)*	*(ii)*	*(i)*	*(ii)*	*(i)*	*(ii)*
3	Drop-out	−1.11	−1.11	1.15	1.15	1.14	1.15
3	3	−1.10	−0.92	1.13	0.56	1.11	0.92
−3	Drop-out	−1.07	−0.89	1.08	1.03	1.10	0.77
−3	−3	−1.10	−1.10	1.08	1.08	1.14	1.14

going on between the statistician and the investigator leading to an iteration between the model and the data.

The statistician's job is to be in close touch with the scientist that he advises and to discuss regularly with him the things that he is finding. For example, the statistician might find out that ailing cows appeared to produce two different kinds of data pattern. The scientist might then supply the information that there were two different kinds of mastitis. Co-operation might then produce two different kinds of model which might describe the behaviour of the two different kinds of disease, and so forth. The conversation would go to and fro and the interchange would greatly affect the analyses that were conducted and the design of further experiments. This is the way that science works. When scientists sometimes dismiss statisticians as being useless to them, I think that it is because we do not interact with them; we separate ourselves off and talk to ourselves.

A. Heyting and J. T. B. M. Tolboom (Solvay Duphar, Weesp): Analysts faced with data suspected to be generated under an informative drop-out (ID) process will certainly welcome the methodology proposed by Diggle and Kenward. However, when planning investigations prone to drop-out, it remains important to *design for ignorability*. This involves measures to contain the number of drop-outs, scheduling frequent measurement times and, if possible, arranging for drop-out times to coincide with measurement times. Above all, the recording of all base-line and treatment emergent drop-out predictors must be called for (see Heyting and Tolboom (1993)).

Achieving a well-fitting and persuasive random drop-out (RD) model for the p_k yields several benefits. Not only does it provide alternative estimates of (θ, ϕ) for comparison, it also gives insight into the drop-out mechanism (for examples see Heyting *et al.* (1990, 1992)). Such insight is crucial for the improvement of future study designs. Most important is that fitting an RD mechanism permits the consistent estimation of the distribution of $\mathbf{Y}^*$ and thus a check on the crucial, but often highly speculative, choice of $f^*(y; \theta, \phi)$ for the fitting of an ID process to data of the current or of a future investigation (for details see the appendix in Heyting *et al.* (1992)). Finally, fitting an RD process is less prone to give rise to a likelihood surface that is flat at the top, at least to a second-degree approximation. Although Diggle and Kenward did not encounter this problem, it is a realistic possibility with an ID process. As a simple illustration, observe that, with univariate $\mathbf{Y}^*$ assumed $N(\mu, \sigma^2)$ and an ID process modelled as $\ln(1-p) = \alpha y^{*2} + \beta y^* + \gamma \leqslant 0$, the observed data are normally distributed (p is the drop-out probability). Fixing $E(p)$ and the distribution of the observed measurements, the model is seen to permit a range of values of (μ, σ^2), including those for the completely RD process. Over relevant parameter ranges, this model is closely approximated by the commonly used linear probit and logit models. Simulations have confirmed that these models yield nearly singular information matrices in some realistic situations.

W. R. Gilks (Medical Research Council Biostatistics Unit, Cambridge): I would like to cast the authors' model in the form of a directed graph. My notation in Fig. 4 follows the authors', except that I_{ij} denotes whether or not the jth measurement for individual i was observed and $D(i)$ indexes the first missing datum. Squares denote observed data; circles denote missing data or parameters.

A directed graph for the authors' model for informative drop-out appears in Fig. 4(a). For random drop-out, the edge (arrow) from the unobserved $Y_{i,D(i)}$ to $I_{i,D(i)}$ would be omitted. The structure of this model is complicated through the tangle of edges in the centre of the graph. Are all these edges necessary and how should the functional part of the model (distributions, etc.) be defined? The answers should come from considering how these dependences might have arisen.

One explanation might be that the dependences are due to latent variables. These are represented by the random effects ρ_i in Fig. 4(b), in the spirit of Wu and Carroll (1988). Conditioning on the random effects, the data are assumed independent, save possibly for a first-order autoregressive process in the Ys (as in equation (16) in the paper). For example, the Ys might represent blood protein measurements, and the ρs the true extent of disease in the liver. Drop-out (death) would be due to the extent of liver disease, not to the biochemical measurements themselves.

Integrating out the random effects in Fig. 4(b) would lead to Fig. 4(a). Thus with suitable functional assumptions the two models are equivalent. The model can in principle be estimated straightforwardly via Markov chain Monte Carlo (MCMC) methods, and for this it would be much simpler to deal directly with Fig. 4(b). Moreover, it is simple to extend the methodology to accommodate measurement error (Richardson and Gilks, 1993) and non-normal errors (Dellaportas and Smith, 1993). An extension of particular interest is to include covariates X, where $X_{i,D(i)}$ is unobserved. It is not clear whether the

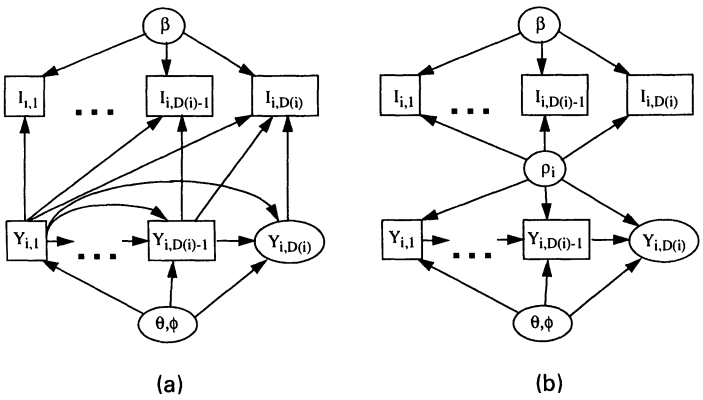

Fig. 4. Graphical models for informative drop-out: (a) Diggle and Kenward's model; (b) a latent variables model

authors' approach could be extended to handle this, but again in principle this is straightforward by using MCMC methods.

A. P. Dawid (University College London): In considering the potentials and the pitfalls in the approach presented, I think—as always—that it is helpful to look at an extreme example.

Let us take a discrete model: coin tossing. To make the model as simple as possible, suppose that we have a number of identical coins. Moreover, there is independence, so there is only a single underlying parameter p (the unknown probability of heads (H)) in which we are interested. The following sequences of outcomes from tossing three coins have been reported: HH, HHHH, HHH; we want to make inferences about p in the possible presence of informative drop-out. We do not quite know why these reported sequences end when they do.

I think that Professor Diggle said that it was dependence in the data sequences that made these problems important, but I do not think that this is true. What would we infer looking at those sequences of data? I would infer that the drop-out process tends to censor tails; in other words, if a tail is about to come up, then drop-out is likely. This suggests that the 'face value likelihood' (Dawid and Dickey, 1977), p^9, ought to be adjusted downwards.

One possibility is to infer that, in each sequence, a tail was the next outcome, but it was censored. This would introduce a further likelihood factor $(1-p)^3$. However, we cannot be sure that a tail would be observed. What I think we learn is that the probability of a drop-out given the current outcome is a tail, $P(D|T)$, is high—it may or may not be 1. We do not, however, learn too much about the probability of a drop-out given that the outcome is a head, $P(D|H)$. In any event information about these quantities will be affected by information about p, and vice versa: there is an intrinsic confounding which may be disguised, but not removed, by imposing strong assumptions.

The two points to learn from this are that

(a) even with independence processes, it matters whether or not the drop-out is informative and
(b) even with a fully specified model—a very simple model—there are aspects of the drop-out process about which it will be very difficult to learn.

Brian Cullis (NSW Agriculture, Wagga Wagga): The result that an increase in drop-out probability occurs for cows with a relatively low level of milk protein is biologically unlikely. The length of a cow's lactation is primarily set by human intervention. Milk yield often declines towards the end of lactation or the cow is rested to prepare for the next calving. Milk yield decline was unlikely for these data given the length of the observation period and the higher plane of nutrition that most cows received.

The data were not originally obtained from the experimenter, but they were presented at a workshop at Adelaide University in 1989. Dr A. Verbyla and I have since contacted Mr S. Valentine, who conducted the experiment, to determine the real reason for the apparent drop-outs. Cows entered the trial as they calved. The experiment was terminated when feed availability declined in the paddock in which the animals were grazing. Thus there are actually no drop-outs as such but five cohorts representing the different starting times.

This incident highlights the dangers of statisticians accepting data at face value without seeking the advice of the experimenters.

We adapt the model proposed by Verbyla and Cullis (1992) for each cohort, i.e. the conditional distribution of $y_i = \text{vec}(Y_i)$, the $n_i p_i$ data vector for the ith cohort, is

$$E(y_i|\beta_i) = X_i\tau + F_i\phi_i + Z_i\beta_i,$$
$$\text{var}(y_i|\beta_i) = \sigma^2(\Omega_{2i} \otimes I_{n_i}) = \sigma^2 V_i$$

and so unconditionally y_i is normal with

$$E(y_i) = X_i\tau + F_i\phi_i,$$
$$\text{var}(y_i) = \sigma^2(\Omega_{2i} \otimes \Omega_{1i}),$$

where $\Omega_{1i} = I_{n_i} + \lambda J_{n_i} J_{n_i}'$, J_{n_i} is the n_i-vector of 1s, $\tau = \text{vec}(T)$ is the vector of treatment effects at each time, ϕ_i is the vector of missing value estimates and β_i is the vector of the ith cohort effects. Estimation proceeds by using residual maximum likelihood (REML) (Verbyla and Cullis, 1992).

The results of the authors' reanalysis suggest cohort effects and this is confirmed in our analysis. For consistency with the authors' analysis we use the same variance model. The variance parameters are estimated from the saturated treatment model. The differences between this and the authors' approach is minor for this data set. However, Cullis and Verbyla (unpublished) have shown that jointly modelling the profile and covariance can lead to losses in efficiency. The REML estimates of the variance parameters were $\hat{\lambda} = 0.416$, $\hat{\sigma}^2\hat{\gamma}_2 = 0.0213$; compare $\hat{\tau}^2$, $\hat{\gamma}_1 = \hat{\rho} = 0.882$ and $\hat{\sigma}^2 = 0.0557$.

Fig. 5 presents the profile for the barley diet and the differences between the mixed and the lupin diet and the barley diet. There is no evidence of a rise in the milk protein towards the end of the experiment, but strong evidence of a decline in milk protein for both the mixed and the lupin only diets relative to the barley diet. The authors' profile model is almost equivalent to the model that both contrast profiles are constant. The test of constancy *versus* a linear decline resulted in $F = 2.59$ ($p = 0.074$).

Donald B. Rubin (Harvard University, Cambridge): This contribution by Diggle and Kenward is an interesting addition to the applied literature dealing with missing values, especially to that concerning the special case of longitudinal data with drop-outs. Several comments may be helpful for relating this work to other contributions, both theoretical and applied.

The general formulation for likelihood-based inference when confronted with missing values (Rubin, 1976) has two conditions for the ignorability of the missing data mechanism: the missing data should be missing at random and the missingness parameter should be distinct from the parameter of the data. The second condition, only briefly mentioned by Diggle and Kenward, has been missed by some biomedical researchers, as noted by Shih (1992), but can be relevant for longitudinal data with hierarchical models where the difference between missing data and random parameters is not always sharp, at least computationally (e.g. Rubin (1992)). Also relevant is the fact that this general formulation can be extended to 'coarse data' situations (Heitjan and Rubin, 1991), which has biomedical applications (Heitjan, 1993).

With ignorable mechanisms, fixed data and a fixed model for the data, likelihood-based inferences for the parameter of the data are unaffected by the specific model for the ignorable mechanism, although even then likelihood-based inferences for estimands also involving missingness parameters (e.g. means for those subjects who do not drop out) generally vary with the posited missingness specification unless it is a completely random one; I mention this because I found the authors' discussion of this basic point rather indirect.

With non-ignorable mechanisms, even inferences for the data parameters generally depend on the posited missingness mechanism, a fact that typically implies greatly increased sensitivity of inference to reasonable model specifications. Moreover, without distributional assumptions (e.g. symmetry or normality of residuals coupled with *a priori* zero regression coefficients as in Greenlees *et al.* (1982), absence of higher order interactions as in Baker and Laird (1988) and the closely related special assumptions imposed by Diggle and Kenward) or supplemental information (e.g. via follow-up surveys on the non-respondents, as in Glynn *et al.* (1993a)), it is impossible to 'test' or find evidence for or against non-ignorability (informative drop-out (ID)); this is an important point that easily might be missed when reading the authors' analyses using their particular ID models. Recent work on non-ignorable missingness with biomedical data includes Little (1993) on pattern mixture models and Lavori (1992) on the need to limit sensitivity through more conscientious data collection efforts. Related work using

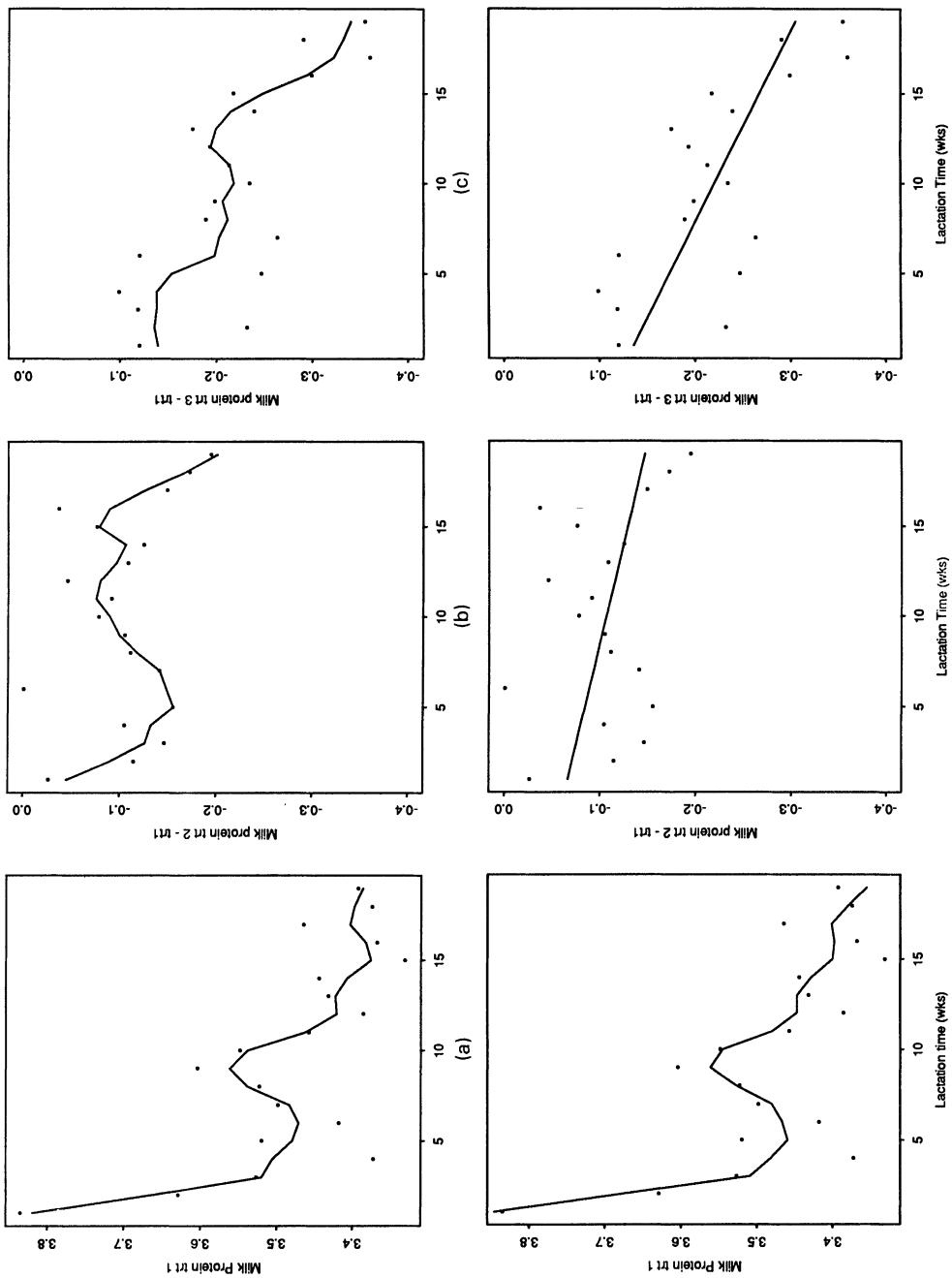

Fig. 5. Profiles for the barley diet and two contrasts: smoothed lowess profiles superimposed for (a)–(d); fitted linear profiles superimposed for (e) and (f)

multiple imputation in biomedical applications, which thereby allows standard complete data methods of analysis, includes Rubin and Schenker (1991) and Tu *et al.* (1993).

The following contributions were received in writing after the meeting.

O. B. Allen (University of Guelph): The authors are to be congratulated on producing an excellent paper on the very important topic of informative drop-out in longitudinal series. I have two queries. The authors make the 'crucial assumption . . . that if an experimental unit is still in the study at time t_k its associated sequence of measurements $\{Y_j: j=1, \ldots, k\}$ follows the same joint distribution as that of the corresponding $\{Y_j^*: j=1, \ldots, k\}$. It is unclear to me what this assumption means. Can the authors provide an example where this assumption would not be reasonable?

My second query is concerned with the milk protein trial. Neither the paper nor the references cited state the point in the lactation at which the 19 weeks of protein percentages were taken. A complete lactation, at least for Holstein–Friesian cows in North America, is considered to be about 43 weeks. If these 19 weeks of data were taken near the beginning of the lactation, the drop-out rate is unusually high. Also, it begins precipitously at week 15 (from the beginning of the trial). We are not told what management decision rules were used to determine when a cow would cease to be milked. Usually, this decision is based on the volume of milk produced, on any health problems and on whether the cow has been successfully rebred. Have the authors considered allowing the probability of drop-out to depend on other covariates in this example? These covariates would typically be correlated with the response vector. Suppose that y_d^* affects the drop-out probability only through another covariate (i.e. the significance of y_d^* vanishes when the covariate is included). Are drop-outs still informative in this case?

Per Kragh Andersen (University of Copenhagen): I shall relate Diggle and Kenward's discussion to special problems encountered in clinical trials in chronic diseases where survival is the main outcome but where repeated recordings of some variable Y may serve as an important secondary end point on which the treatment effect is to be evaluated (e.g. Lombard *et al.* (1993)). In such a study survival may be modelled via, for example, a Cox regression model where the hazard function for patient i at time t is given by

$$\alpha_i(t) = \alpha_0(t) \exp\{\beta^T Z_i(t)\} \tag{34}$$

and where the vector of covariates $Z_i(t)$ includes (from the history $H_i(t)$ for patient i at time $t-$) a treatment indicator and information on recordings of Y_i in $[0, t)$. For *discrete* Y, joint modelling of survival and the development of Y may, in some cases, be performed within a Markov process framework (Kay, 1986; Andersen *et al.*, 1991) but for *quantitative* Y no such methods are available. However, the present paper provides a framework for joint modelling, at least when the survival times T_i are grouped into intervals $[t_{k-1}, t_k)$ independent of i. In this case model (34) implies that

$$P\{T_i \leqslant t_k | T_i > t_{k-1}, H_i(t_k); \beta\} = 1 - \exp[-\exp\{\beta_0^{(k)} + \beta^T Z_i(t_k)\}] \tag{35}$$

with

$$\beta_0^{(k)} = \log\left\{\int_{t_{k-1}}^{t_k} \alpha_0(t)\, dt\right\},$$

and assuming that no new observations of Y_i are available in (t_{k-1}, t_k). Model (35) corresponds to model (18) with

(a) $\beta_1 = 0$ (no informative drop-outs),
(b) β_0 replaced by $\beta_0^{(k)}$ and
(c) the logit link function replaced by $\log\{-\log(1-p)\}$.

The assumption of no informative drop-outs is necessary to ensure that model (34) may be interpreted as a hazard function, i.e. that patient i's risk of failing at time t only depends on information in $[0, t)$ and not, for instance, on the value of Y_i that would have been obtained at time t had patient i been alive at that time. The paper then both provides a means of checking this assumption and a method for joint analysis of Y and T. Under non-informative drop-out this analysis further reduces to a separate analysis of the development of Y based on, for example, equations (15) and (16) and an analysis of a (grouped) Cox regression model for T with previous recordings of Y as time-dependent covariates.

In survival studies there will always be *right censoring* and a crucial point is then whether this censoring may be assumed to be *independent*, i.e. whether the presence of censoring invalidates the inference on

the survival times. A similar question may be raised for a quantitative response variable Y measured over time. Thus Wu and Bailey (1988) (see also Wu and Carroll (1988)) studied a random slopes model where the distribution of a censoring time may depend on this slope. The present paper provides an alternative where the censoring time may depend in a more general way on previous measurements of Y as well as on the potential future measurement.

Kevin Gough (Fisons plc, Loughborough): Drop-outs are a major concern in the design and analysis of clinical trials in that they are a potential source of bias in treatment comparisons. Methods used are often *ad hoc*, e.g. carrying forwards the last observed value, and this paper is welcome as it offers a rigorous method of dealing with drop-outs.

When patients drop out of a clinical trial the reason for withdrawal is often recorded. Might it not be reasonable to fit different drop-out processes for different types of drop-out? For example in the antidepressant trial the drop-outs could be divided into those who dropped out for lack of efficacy and those who dropped out for other reasons (e.g. adverse events, protocol violations, etc.). For the lack of efficacy drop-outs the process could be modelled as in the paper; however, for the other drop-outs a random drop-out or completely random drop-out model might be sufficient. In the paper it was noted that evidence for an informative drop-out process for all drop-outs was only borderline; by focusing on the lack of efficacy drop-outs the evidence may become more clear-cut.

Richard Kay (S-Cubed, Sheffield): Earlier Matthews raised questions concerning 'intention to treat' in the analysis of clinical trials. My question concerns the way that such analyses are often achieved in practice. Usually when a patient drops out of a trial their last observation is used as the response in a 'last observation carried forward' analysis. Clinically this is a sensible approach in that this value records what we have been able to achieve with that patient. From a modelling standpoint, however, it is difficult to see precisely which underlying process we are addressing. It is not $\mathbf{Y}$ nor $\mathbf{Y}^*$. I would value the authors' comments.

Secondly, there are connections between this work and concepts of informative and non-informative censoring in survival analysis (Gruger *et al.*, 1991). Censoring is non-informative provided that the probability that a patient is censored depends only on the past and not on the future. With informative censoring it is not possible to make inferences on the underlying survival time distribution, only on the combined survival–censoring process as it pertains.

Finally, returning to longitudinal data I was somewhat surprised that no comments were made on design. We very much tend to concentrate on developing sophisticated methodology to cope with problem data. Many 'missing' data problems can be avoided, or at least minimized, through better design and more careful attention to the conduct of our trials and the data collection process. We should never lose sight of those things.

P. J. Kelly (University of Newcastle upon Tyne): This paper is an important contribution to a class of problem that is particularly widespread in medicine. Clinical trials into the investigation of pain relief given by various drug preparations are common in medical research. In particular, there are many trials that assess acute post-operative pain by using a series of 10 cm visual analogue scores measured at each of the times $t_1, \ldots, t_n$. The three examples discussed in this paper are all long-term studies whereas acute pain studies typically consist of a short sequence of half-hourly or hourly measurements. The majority of such trials offer the patient an escape analgesia which the individual is allowed to take if greater pain relief than that given by the trial medication is required. If patients do take an escape then they are considered to have dropped out of the trial. A search of the journals *Anaesthesia* and *Anesthesia and Analgesia* for 1991 reveals 21 such papers where the group sizes, number of time points and the numbers taking escapes are described (there being a number of others without this information).

TABLE 9

	Median	*Interquartile range*	*Range*
Size of treatment group	20.5	15.0–25.0	8–50
No. of time points	7	4.5–8.0	3–15
No. taking escape as % of no. in trial	47.0	28.0–68.5	10.0–87.0

There were 13 trials with two treatment groups and four each with three and four groups respectively. Table 9 summarizes some relevant information.

Several questions arise.

(a) How will the methods proposed in the paper cope with a 'typical' trial as represented by the medians shown in Table 9?
(b) Some studies record the time of a drop-out exactly and also y_{ij} at that time; they may or may not continue to record further responses at later times. Others only record that a drop-out occurred in an interval and only have information on the response variable up to the previous time point. Ideally, should such trials be designed with the aim of yielding random drop-outs or informative drop-outs and is it always possible to distinguish them from each other?
(c) What effect will the relatively short time spans of such trials have on the analysis?

Nan M. Laird (Harvard School of Public Health, Boston): Although the term informative drop-out (ID) is now in increasingly popular use, its precise meaning is often misunderstood. Papers such as this are helpful in so far as defining specific models aids in clarifying terminology.

I view the use of these 'non-ignorable' models with much scepticism for two reasons. First, estimating the 'unestimable' can be accomplished only by making modelling assumptions, either parametric distributional assumptions or assumptions on associations. This can be clearly seen in dealing with discrete responses (see, for example, Baker and Laird (1988)). With measured responses and multiple variables it is much more difficult to understand the model limitations and hence sensitivity to assumptions. We do not have a clear understanding of what the consequences of model misspecification are, even in the ignorable case. The consequences of model misspecification will probably be far more severe in the non-ignorable case. See, for example, the discussion in Glynn *et al.* (1993b) and references cited therein. Sensitivity to model assumptions is well illustrated in the authors' examples where the results for the ID models depend strongly on whether or not $\beta_2 = 0$ (see $\hat{\eta}$ from Table 4 and $\hat{\delta}$ from Table 5). I am not very comforted by the use of likelihood ratio tests for distinguishing between models, since different models may have different hidden assumptions.

My second view of these models is that they may be more useful at raising questions than they are in providing answers. The apparent lesson learned from the use of these models in example 1 is that the observed increase in response profile is due to failure to account for drop-out. Yet this could easily be learned from looking only at the random drop-out (RD) or completely random drop-out (CRD) models. (Is there a typographical error in Table 4? The results for RD and CRD should be identical but differ in the second decimal for $\hat{\eta}$.) If you fit an ordinary least squares (OLS) model and use the sandwich variance estimate you *might* obtain a misleading inference about η or ξ. Using OLS with a variance correction is really more in the spirit of what the authors mean to imply by a CRD drop-out analysis.

In the second example, the focus of inference is on the drop-out process, not milk yields. If one is primarily interested in estimating the average increase in milk yields between years 3 and 4, what is the answer? Is it around 0.3 or 0.7? How is the inconsistency in the two ID models explained?

The third example illustrates the point raised in Wu and Bailey (1988), Little (1983) and discussed in Glynn *et al.* (1993b), namely that, if the drop-out model assumes no treatment–Y_k interaction, then inferences about treatment effects are largely unaffected by the use of non-ignorable models.

E. Lesaffre and G. Molenberghs (Limburgs Universitair Centrum, Diepenbeek): Because of a lack of flexible models incorporating missing data mechanisms, drop-outs are not treated (adequately) in many trials. We welcome this work as a major step in the development of methods filling this gap.

Defining the full data as (Y_{obs}, Y_{mis}, R), with R a missing data indicator matrix, facilitates model formulation and construction of fitting algorithms (Little and Rubin, 1987). For drop-outs, R can be defined as $R_i = j$ if subject i drops out at t_j, $R_i = n_i + 1$ otherwise. A parsimonious description results, from which appropriate distributions are easily derived. The less standard notation (Y^*, Y) immediately follows.

The EM algorithm (Dempster *et al.*, 1977) provides a convenient maximization technique for a broad class of problems. It converges under mild conditions, although at the cost of linear convergence. At the final expectation step, imputed values for Y_{mis} are obtained. Did the authors choose the simplex method for any particular reason?

We want to reiterate the important point of considering various plausible models for the drop-out mechanism, as a way of assessing the influence of misspecification.

To have a precise idea of the drop-out process it is important to observe a minimal proportion of

drop-outs, without reducing the information too much. The examples and the simulations give a rough idea. Could the authors elaborate on this?

Formulae (18) and (20) provide an appealing general framework to model drop-out. We want to exemplify the potential importance of covariates. Consider a clinical trial where the therapeutic effect is of primary interest but the severity of side-effects affects drop-out. Patients with similar observed (early) responses but with different levels of side-effects would have different drop-out probabilities and/or different unobserved responses. To assess this difference, the inclusion of covariates in the drop-out process can be crucial. However, with many covariates available (see model (20)), it could be difficult to distinguish between random drop-out (RD) and informative drop-out (ID); see also Laird (1988).

The more powerful models incorporating drop-outs become, the sloppier experimental researches could become, supposing that statisticians could handle the resulting messy data sets. But every single drop-out results in a loss of statistical information. Can the model building process partly recover the variability by considering ID under an ID assumption, compared with considering RD under ID?

As Lavori (1992) states, 'It is always better to have no dropouts': an evident, yet important, statement.

Roderick J. A. Little (University of Michigan, Ann Arbor): This paper has many strengths, but the potentially serious lack of robustness of the authors' informative drop-out (ID) models needs more emphasis. These models extend the logit selection model for univariate missing data of Greenlees *et al.* (1982). Very similar probit selection models (Heckman, 1976; Amemiya, 1984) are widely used in some areas of econometrics, but they have been attacked on the grounds that estimates based on them are highly sensitive to untestable distributional assumptions (Little (1982, 1985), Glynn *et al.* (1986), Tukey (1986) and Little and Rubin (1987), chapter 11). Statisticians adopting similar ID models for longitudinal attrition should consider these references carefully.

Consider a single drop-out time, and let Y_1 denote the (fully observed) variables up to drop-out and Y_2 the (incompletely observed) variables after drop-out. The data clearly supply no direct information about the distribution of Y_2 given Y_1 for subjects who drop out. Under RD, this distribution is assumed the same as that for subjects who do not drop out. Under ID, differences in the distribution of Y_2 given Y_1 for those who do and do not drop out are solely determined by distributional assumptions of the model, such as the form of the model for drop-outs, normality or constraints on the mean and covariance matrix. The authors' simulations show (predictably) that bias can be reduced if the model is correctly specified but provide no evidence of stability or robustness to misspecification.

An alternative approach is pattern mixture modelling (Little, 1993a; Glynn *et al.*, 1986), where the sample is stratified by missing data pattern and each stratum modelled separately. In Little (1993b) for scalar Y_1 and Y_2 and Little and Wang (1993) for vector Y_1 and Y_2, parameters are identified by assuming that missingness is an arbitrary function $g(\beta_2 Y_1 + \beta_1 Y_2)$; the form of $g(\ )$ does not need to be specified. The data provide no information about β_1, so a sensitivity analysis over a range of plausible values of β_1 is proposed. This approach is related to the authors' ID model, with β_1 prespecified rather than estimated from the data (see also Okafor (1982)). A sensitivity analysis is less appealing but may be more realistic in terms of what can be estimated from the data.

K. V. Mardia (University of Leeds): It is nice to see a combination of a multivariate linear model and a logistic regression model. My comments will be mainly restricted to the updating equations (9) and (10) of the paper in the context of universal kriging and thin plate splines. Let $\mathbf{x}_i$, $i = 1, \ldots, k$, be k spatial locations (say in two dimensions) with realization $z_i = z(\mathbf{x}_i)$, $i = 1, \ldots, k$, of a random field. Suppose that $z_{k+1} = z(\mathbf{x}_{k+1})$ is a new observation at a point $\mathbf{x}_{k+1}$. Let $z_k(\mathbf{x})$ be the universal kriging predictor based on k locations, say under a linear trend. Then we can write

$$z_k(\mathbf{x}) = \mathbf{z}_+^{\mathrm{T}} [\, \Sigma_+^{-1} \ \sigma_+(\mathbf{x}) \,]$$

where

$$\sigma_+(\mathbf{x}) = (\sigma(|\mathbf{x} - \mathbf{x}_1|), \ldots, \sigma(|\mathbf{x} - \mathbf{x}_k|), 1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}},$$

$$\mathbf{z}_+^{\mathrm{T}} = (\mathbf{z}^{\mathrm{T}}, \mathbf{0}_3^{\mathrm{T}}), \qquad \mathbf{z}^{\mathrm{T}} = (z_1, \ldots, z_k),$$

$$\Sigma_+ = \begin{pmatrix} \Sigma, & \mathbf{Q} \\ \mathbf{Q}^{\mathrm{T}}, & \mathbf{0} \end{pmatrix}, \quad \mathbf{X}^{\mathrm{T}} = \begin{pmatrix} 1, 1, \ldots, 1 \\ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \end{pmatrix},$$

$$(\Sigma)_{ij} = \sigma(|\mathbf{x}_i - \mathbf{x}_j|) = \sigma_{ij}.$$

Then, on following result (11), the predictor on adding one extra point is simply

$$z_{k+1}(\mathbf{x}) = z_k(\mathbf{x}) + z_k'(\mathbf{x}) - c_{k+1}\,\sigma(|\mathbf{x} - \mathbf{x}_{k+1}|), \tag{36}$$

where

$$z_k'(\mathbf{x}) = c_{k+1}\,\boldsymbol{\sigma}^{\mathrm{T}}_{+(k+1)}\,[\,\boldsymbol{\Sigma}^{-1}_{+}\;\;\boldsymbol{\sigma}_{+}(\mathbf{x})\,]$$

with

$$\boldsymbol{\sigma}^{\mathrm{T}}_{+(k+1)} = (\sigma_{1,k+1}, \ldots, \sigma_{k,k+1};\ 1,\ \mathbf{x}^{\mathrm{T}}_{k+1})$$

and

$$c_{k+1} = -(\boldsymbol{\sigma}^{\mathrm{T}}_{+(k+1)}\,\boldsymbol{\Sigma}^{-1}_{+}\,\mathbf{z}_{+} - z_{k+1})/\boldsymbol{\sigma}^{\mathrm{T}}_{+(k+1)}\,\boldsymbol{\Sigma}^{-1}_{+}\,\boldsymbol{\sigma}_{+(k+1)}.$$

Thus the effect of adding one extra point can be understood through equation (36). In particular, the terms in the square brackets are the same in $z_k(\mathbf{x})$ and $z_k'(\mathbf{x})$. Indeed, equation (36) underlies very recent developments of incorporating derivative information around any specific point $\mathbf{x}_i$ as pioneered by Bookstein and Green (1993) and extended by Mardia and Little (1993). This work allows the incorporation of edge information around landmarks in images. For thin plate splines, $\sigma(r) = r^2 \log r^2$. The updated prediction mean-square error (kriging variance) can be written similarly but the expression is more complicated.

A model of repeated multivariate measurements in the context of a spatial–temporal setting is considered in Mardia and Goodall (1993) which may prove relevant when the longitudinal data are multivariate.

D. A. Preece (University of Kent, Canterbury): Will the authors please state what sort of infection mastitis is, and how the infection is believed to be spread? Without this information, surely no proper assessment can be made of any model proposed for the data from the mastitis study.

Stephen Senn (Ciba–Geigy, Basle): Presented with such a useful and interesting paper which covers so many important points it may seem churlish to ask for more but some extensions and parallels are suggested and it would be interesting to have the authors' opinion.

In this paper the continuous measurements on completers are of primary interest but by modelling the drop-out process further information is recovered from non-completers. In the context of survival analysis the emphasis would be reversed: time to drop-out for non-completers would become the primary outcome but a continuous surrogate might provide useful information from the completers. It might be interesting to explore the connection between this paper, survival analysis and the method proposed by Gould (1980) cited by the authors. Also of interest might be the connection to surrogate end points (Prentice, 1990). A very useful (but no doubt difficult) extension of the paper would be to cover partial drop-out (non-compliance) and the use of rescue medication in clinical trials. What prospects do the authors see here?

It would be useful to have more details regarding the simulations. The values of t_j are undefined for the first example in Section 4. Does $t_j = j - 1$? In the case of the crossover it would be useful to have the row and column means. Without this information it is difficult to judge under which treatments the drop-outs mainly occur. (No clues are given about the values under C and D.) In fact, if space permits it would be nice to have a rough idea of the proportion of drop-outs under a given treatment and/or the values of β_1 and β_2 used in the simulation. I find the argument regarding correction of treatment effects for period effects at the end of Section 4 rather confusing. One 'effect' appears to be a contrast and the other a (non-estimable unless arbitrarily parameterized) mean. Whether period and treatment contrasts are negatively or positively biased must surely depend on the way that they are defined. Intuitively I would have thought that the general effect of any realistic drop-out process would be to reduce in absolute value (attenuate) any treatment contrast but perhaps with crossover trials the position is more complex.

It might also be worth noting that the crossover design is very commonly used in single-dose pharmacodynamic studies. In my experience in such cases it is more common for a patient to drop out and to take rescue medication on a particular day but to return to take subsequent treatment than to drop-out altogether.

Weichung J. Shih (Merck Research Laboratories, Rahway): As advocated in Shih (1992), I am happy to see the corresponding use of random drop-out (RD) and completely random drop-out for longitudinal missing data by the authors to the terms missing at random (MAR) and missing completely at random which have been used for the general missing value problem. However, the term '*informative* drop-out' seems to have been used less uniformly; see for example Wu and Carroll (1988), Wu and Bailey (1989) and Schluchter (1992) for the two-stage random effects model. If we view the individual slope coefficients in the two-stage random effects model as latent variables (see Shih *et al.* (1993)), as opposed to 'parameters' (as in Laird and Ware (1982)), then informative drop-out means the same as violation of RD (or MAR). 'Informativeness' is not equivalent to 'non-ignorability' in Rubin's (1976) original framework for missing data inference, since RD (or MAR) is only one of the conditions for ignorable missing data mechanisms. The other condition, 'distinct parameters' (DPs), also deserves attention (see also Little (1993)). When MAR holds but the DP condition does not, ignoring the missing data mechanism results in loss of efficiency, but not consistency. This is a far reaching result as the following example demonstrates. Consider the pair of binary random variables (x, y) with probability density function $p\{(x, y)=(i, j)\}=\theta_{ij}$, $i, j=0, 1$. Assume that x is always observed and that $p(y=*)=\phi=\theta_{00}^2$, where $*$ denotes a missing value. Data contain n complete pairs, with n_{ij} pairs of (i, j), $\Sigma n_{ij}=n$, and m incomplete pairs, with m_0 pairs of $(0, *)$ and m_1 pairs of $(1, *)$, $m_0+m_1=m$. For simplicity, assume $\theta_{01}=0$ so that only θ_{00} and θ_{10} need to be estimated. The full likelihood based on the observed data and the missing data pattern is

$$L=(1-\phi)^n\,\theta_{00}^{n_{00}}\,\theta_{10}^{n_{10}}\,(1-\theta_{00}-\theta_{10})^{n-n_{00}-n_{10}}\,\phi^m\,\theta_{00}^{m_0}(1-\theta_{00})^{m-m_0},$$

where $\phi=\theta_{00}^2$. We need considerable algebra to obtain the maximum likelihood estimates (MLEs). For example, the MLE of θ_{00} is $\hat{\theta}_{00}=\{-b+\sqrt{(b^2-4ac)}\}/2a$, where $a=3(m+n)$, $b=n+m-n_{00}-m_0$ and $c=-(n_{00}+2m+m_0)$. However, if we ignore the missing data mechanism by recognizing that it is a case of MAR without DP, immediately we have the familiar complete data situation and obtain the usual estimate $\tilde{\theta}_{00}=(n_{00}+m_0)/(n+m)$ based on the proportion of (0, 0) pairs in the total sample. This estimate is not the MLE but is consistent. As the example illustrates, the potential gain in simplicity can be considerably advantageous for some very complicated situations such as the longitudinal data. The next question is, how much is the loss of efficiency? Although Altham (1984) has shed some light on this, more research is needed to answer the question. For the above example, a simulation of 500 runs showed that the relative efficiency of $\tilde{\theta}_{00}$ to $\hat{\theta}_{00}$ is about 0.51.

A. P. Verbyla (University of Adelaide): I congratulate the authors on a very interesting paper. My comment concerns the estimation procedure and was developed in collaboration with Dr Brian Cullis.

For estimation of Gaussian linear models, i.e. model (1), it is generally accepted that residual maximum likelihood (REML) is preferable to maximum likelihood (ML); see Patterson and Thompson (1971). In this case, REML provides a simple adjustment to the profile likelihood for ϕ, namely $\det(X^{\mathrm{T}}V^{-1}X)^{-1/2}$, and as θ and ϕ are orthogonal (assuming that they are functionally independent) this corresponds to modified profile likelihood (Barndorff-Nielsen, 1983) and approximate conditional profile likelihood (Cox and Reid, 1987).

In the current setting, if the process is completely random drop-out or random drop-out, the orthogonality extends to β and REML proceeds as in the standard case. This is not true under informative drop-out (ID), which is not very surprising given the implication of ID. Any approach seems very difficult; for example θ and ϕ are no longer orthogonal. If ϕ and β are treated as nuisance parameters and θ is the parameter of interest, one approach is to adjust the profile likelihood for ϕ and β by the term

$$\det\{j_{\theta\theta}(\hat{\theta}_{\phi,\beta})\}^{-1/2}=\det\{X^{\mathrm{T}}V^{-1}X+\nabla_\theta\beta_0^*\,W^*(\nabla_\theta\beta_0^*)^{\mathrm{T}}\}^{-1/2}$$

where $j_{\theta\theta}$ is the observed information matrix of θ evaluated at $\hat{\theta}_{\phi,\beta}$ the estimate of θ for given ϕ and β, $\nabla_\theta\beta_0^*$ is a matrix of derivatives of $\beta_{0,i}^*$ (one for each drop-out unit) with respect to θ and W^* is a diagonal matrix with elements $p_i^*(1-p_i^*)$, $p_i^*=P(Y_{d_i}=0|H_{d_i}, Y_{d_i-1}\neq 0)$, again for each drop-out unit. The observed information is seen to be the contribution of the Gaussian and logit components in the model. Including this in the likelihood only corresponds to part of the correction of the modified profile likelihood. However, unless there is a large drop-out rate, which in practice would itself suggest potential problems, we conjecture that the standard REML correction will be sufficient. This needs to be examined, although the simulation study by the authors under ML lends some support to this conjecture.

Margaret C. Wu (National Heart, Lung, and Blood Institute, Bethesda): I would like to congratulate the authors for their clear illustrations of the differences between the drop-out models. The following two items are suggested for discussion.

First, in clinical trials and epidemiological studies serial measurements of the response variables are usually subject to measurement error. Participants who are sick tend to miss the follow-up visits or to drop out of the study to seek advice from their own physicians. Furthermore, the probabilities of sickness and hence of drop-out are usually dependent on the underlying changes in the 'true' response. Hence, equation (32) can be viewed as a logistic regression model when covariates are subject to measurement error. In such situations, the usual estimator obtained by regressing on the observed covariates is asymptotically biased. Mori *et al.* (1991) proposed a similar logistic regression drop-out model using the method of Stefanski and Carroll (1985) to account for measurement error. Correcting for measurement error and its effect on parameter estimation of the primary response could be important when the measurement error variance and sample size are such that the bias in the usual methods becomes large.

The second point concerns an alternative approach, the conditional linear model (Wu and Bailey, 1989), to account for informative drop-out. Under this approach the joint distribution $f(\mathbf{Y}_i, \alpha_i, \mathbf{D}_i|\mathbf{X}_i)$ of the response vector $\mathbf{Y}_i$, the random parameters α_i describing the changes over time of the individual responses and the drop-out variables $\mathbf{D}_i$ (e.g. the time of drop-out and the number of observations made or missed) given some fixed covariates $\mathbf{X}_i$ (e.g. treatment groups) is represented as a product of three factors, i.e.

$$f(\mathbf{Y}_i, \alpha_i, \mathbf{D}_i|\mathbf{X}_i) = f_1(\mathbf{Y}_i|\alpha_i, \mathbf{D}_i, \mathbf{X}_i)\, f_2(\alpha_i|\mathbf{D}_i, \mathbf{X}_i)\, f_3(\mathbf{D}_i|\mathbf{X}_i),$$

where $f_1(\mathbf{Y}_i|\alpha_i, \mathbf{D}_i, \mathbf{X}_i) = f_1(\mathbf{Y}_i|\alpha_i, \mathbf{X}_i)$ models the within-individual response, $f_2(\alpha_i|\mathbf{D}_i, \mathbf{X}_i)$ models the conditional distribution of the within-individual changes given the drop-out patterns and $f_3(\mathbf{D}_i|\mathbf{X}_i)$ models the marginal distribution of the drop-out patterns. This approach is similar in spirit to the pattern mixture models of Little (1993) and the stratified summary statistic approach of Dawson and Lagakos (1991).

Follmann and Wu (1992) showed that the distribution of $(\alpha_i|\mathbf{D}_i)$ is stochastically ordered with respect to any drop-out variable $\mathbf{D}_i$, provided that the distribution of $(\mathbf{D}_i|\alpha_i)$ is totally positive of order 2 (this includes the exponential family). It follows that $E[\alpha_i|\mathbf{D}_i]$ is monotone and can be approximated by using monotonic functions of $\mathbf{D}_i$ for a large class of drop-out models. In most practical situations a linear function of $\mathbf{D}_i$ could account for most of the information. Simulations indicated that the conditional linear model estimator with bootstrap variance is robust under moderate and realistic model deviations (Wu *et al.*, 1993).

The **authors** replied later, in writing, as follows.

We thank all the discussants for the interest which they have shown in our work, and for their constructive criticism of our ideas. We shall first comment on a number of recurrent themes in the discussion, then respond to some individual points.

Dawid, Laird, Little, Matthews, Rubin and Skinner all commented on the sensitivity of our analyses to model assumptions which are difficult to verify from the available data. We agree, and suspect that the same must be true of *any* attempt to make inferences in the presence of informative drop-outs (IDs), unless considerations external to the data determine the drop-out model (as, for example, in classical censoring of observations outside well-defined limits—like Laird, we feel that it is misleading to use 'censoring' as a synonym for 'drop-out'). However, this suggests that more attention should be paid to sensitivity analyses under different plausible assumptions about the drop-out process.

The remarks of many of the discussants (Allen, Andersen, Gilks, Gough, Grieve, Henderson, Kay, Little, Matthews, Senn and Wu) provide ample scope for different modelling approaches, and we apologize that we have space to respond only to some of these. Grieve draws attention to the possibility that drop-outs may form a subpopulation with an identifiably different pattern of responses well before the actual time of drop-out. This suggests modelling the drop-out probabilities in terms of unobserved variables such as the random intercept U in our equation (15), rather than on the measurement process Y^*. Wu and Gilks advocate models of this kind. Gilks also points out that such models produce a simpler graphical structure than our measurement-driven models. We agree that our models appear complicated *when viewed as graphical models* but fail to see why this provides evidence for or against their value as models for actual data. The details of Gilks's proposed model, such as a first-order autoregressive

component which is not particularly natural for observations in continuous time, seem to be motivated primarily by a desire to achieve a simple graphical structure as an end in itself, and in any case this simplicity is lost when a measurement error component is added to the model. Nevertheless, we are pleased to learn that the associated inferential machinery can cope with measurement error, and we agree that the removal (in principle) of the restriction to Gaussian models for the measurement process is an important advance. Andersen, Henderson, Kay and Senn point out possible connections with proportional hazards modelling of survival data. We are intrigued by these and would like to explore them in more detail. One particular aspect mentioned by Andersen, and by Grieve, is the need to understand better what is happening between t_{d-1} and t_d. Here, and in other aspects of longitudinal data analysis, we feel that the underlying continuous time setting should influence the model formulation. This is an argument against both conditional independence modelling based on a fixed set of observation times and our own logistic regression model for drop-outs.

Cullis, Dawid, Grieve, Kay, Rubin, Shih, Skinner and Verbyla all touch on aspects of inference. Firstly, we agree that residual maximum likelihood (REML) is preferable to ML for parameter estimation, but note from Verbyla's comments that this is far from straightforward. It is not even entirely clear what is the appropriate definition of REML for non-linear models, since there are several different principles which can be invoked to derive the REML estimators in the familiar linear Gaussian setting but which give different answers in the non-linear, non-Gaussian case. We accept the criticism of several discussants who disliked our emphasis on point estimates and likelihood ratio tests. As things stand, our software implementation is simply too crude to permit the routine exploration of likelihood surfaces which we would very much prefer.

These unresolved problems of model formulation and inference underline the importance of the remarks by Heyting and Tolboom, by Kay, and by Lesaffre and Molenberghs on the design of studies to avoid IDs whenever possible. However, we would like to emphasize our view that the serial correlation inherent in longitudinal data makes interpretation a subtle business even in the random drop-out (RD) case. Incidentally, we agree (of course) with Dawid that *informative* drop-outs affect the interpretation of correlated *or* uncorrelated data, and apologize if our presentation at the meeting suggested otherwise—the example which we showed at the meeting, taken from chapter 11 of Diggle *et al.* (1994), concerned the contrast between the behaviour of an RD model with correlated and with uncorrelated measurement processes, and was intended to amplify our remarks in the last two paragraphs of Section 2 of the paper.

Matthews and Skinner suggest that the bias in RD estimation when the true drop-out process is ID may be more than compensated by a smaller variance, so that RD estimators may give a smaller mean-square error (MSE) than the strictly correct ID estimators. Maybe so, but the real issue is inconsistency rather than bias itself (we accept the criticism that our *presentation* of the simulation results overemphasized bias). Asking whether in general there is much information on ID is like asking how long is a piece of string. It depends on the problem, sample size and model. Clearly, for sufficiently large sample sizes the bias will dominate the variance in MSE calculations, and vice versa. In our ID simulations, the differences in MSE between the RD and ID estimates were of the order of 5% at most and commonly much less. For these reasons, we would feel very uncomfortable in recommending an RD analysis when the drop-out process is known to be ID, without establishing that the results would be essentially unchanged. In this context, we agree with Laird that our models are likely to prove more useful in raising questions than in answering them, but we certainly regard this as worthwhile since it forces the scientist to address interpretational questions which might otherwise be swept under the carpet. In the milk yield example, if pressed we would plump for a point estimate $\hat{\delta} = 0.3$, because the value $\hat{\delta} = 0.7$ is obtained from a model which is demonstrably inconsistent with the data. But (to hark back to an earlier point), we would prefer to develop our implementation to enable routine calculation of a confidence interval for δ based on our preferred model.

Preece asks for more information about mastitis in dairy cattle. Mastitis is an inflammation of the mammary gland that is characterized by physical and chemical changes of the milk. It has a complex aetiology, involving bacteriological agents. Many factors contribute to the spread of the disease, husbandry appearing to be one of the key components. Preventative measures include regular testing and maintenance of milking machines, care with hygiene and post-lactation use of antibiotics.

Kelly suggests an interesting area of application in which we would expect that IDs would be the rule rather than the exception. Also, the short time span of the trials which he describes is likely to lead to strongly autocorrelated response sequences, which in our view makes it all the more important to model the drop-out process in order to interpret the results.

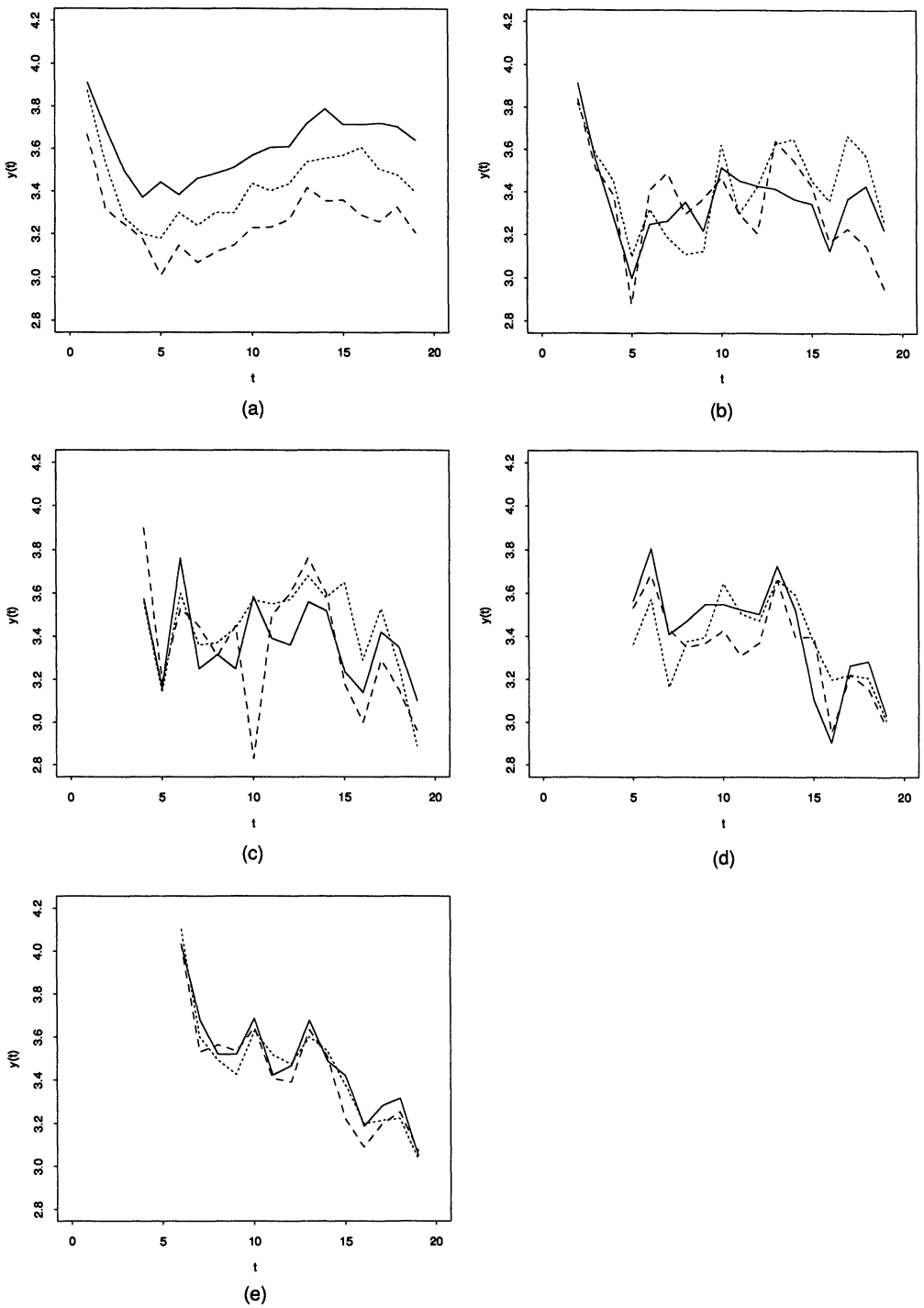

Fig. 6. Observed mean response profiles for the milk protein data, by treatment and cohort (treatments are barley (———), mixed (---------) and lupins (– – – –); time is in calendar weeks, with the start of the experiment as week 1): (a) cohort 1; (b) cohort 2; (c) cohort 3; (d) cohort 4; (e) cohort 5

Mardia reminds us that models with correlation structure similar to our model for the measurement process are also used in spatial statistics. One of the pleasing aspects of recent statistical research is how formerly separate compartments have continued to merge, so that nowadays multivariate methodologists, time series types and spatial statisticians can all agree that they are in the common business of interpreting dependent data structures.

Senn points out that intermittent missing values, which are specifically excluded in our modelling framework, can also be informative, especially in the area of clinical trials where compliance is voluntary. This needs more investigation.

We are grateful to Cullis for putting the record straight on the milk protein data. As he says, the earlier published analyses by Verbyla and Cullis (1990) and by Diggle (1990) must now be viewed as analyses on a biological timescale starting at calving. Curiously, the recognition of cohort effects in these data seems to make their interpretation very much more difficult than before. Cullis models cohort effects as additive nuisance parameters, but our re-examination of the data suggests that the cohort effects are much more complicated than this, as shown in Fig. 6. Note in particular that in cohort 1, which accounts for 41 of the 79 animals, there is clear evidence of a 'settling-in period', with a decrease in mean response over the first 3 weeks before a gentle increase over the remainder of the experiment, whereas in cohort 5, which has 20 animals, the mean response decreases steadily throughout. The pattern is less clear in the intermediate cohorts, where the replication is very low (5, 4 and 9 animals respectively).

Finally, Box criticizes us for failing to interact with our scientists. We agree that it is vital for statistics *as a profession* to interact in this way, and we have both done so on many occasions. But we also believe that it is important for *some members of our profession* to step back from the consulting front line and to develop new methodological tools which can allow future consultants to do a better job. We take some satisfaction from the fact that it was our analysis of the milk protein data, conducted in the spirit of statistical research, which revealed the true story—albeit with the help of Cullis's first-class statistical detective work! If our discipline is to continue to thrive, we must have room for research specialists as well as general applied consultants.

References in the Discussion

Altham, P. M. E. (1984) Improving the precision of estimation by fitting a model. *J. R. Statist. Soc.* B, **46**, 118–119.

Amemiya, T. (1984) Tobit models: a survey. *J. Econometr.*, **24**, 3–61.

Andersen, P. K., Hansen, L. S. and Keiding, N. (1991) Non- and semiparametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process. *Scand. J. Statist.*, **18**, 153–167.

Baker, S. and Laird, N. M. (1988) Regression analysis with categorical data subject to nonignorable nonresponse. *J. Am. Statist. Ass.*, **83**, 62–69.

Barndorff-Nielsen, O. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Bookstein, F. and Green, W. D. K. (1993) A feature space for edgels in images with landmarks. *J. Math. Imgng Vis.*, **3**, 231–261.

Brumby, S. (1989) Exchange of comments on the simplex algorithm culminating in quadratic convergence and error estimation. *Anal. Chem.*, **61**, 1783–1786.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.

Dawid, A. P. and Dickey, J. M. (1977) Likelihood and Bayesian inference from selectively reported data. *J. Am. Statist. Ass.*, **72**, 845–850.

Dawson, J. D. and Lagakos, S. W. (1991) Analyzing laboratory marker changes in AIDS clinical trials. *J. Acq. Immune Def. Synd.*, **4**, 667–676.

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.*, **42**, 443–459.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Diggle, P. J. (1990) *Time Series: a Biostatistical Introduction*. Oxford: Oxford University Press.

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

Follmann, D. and Wu, M. C. (1992) An approximate generalized linear model with random effects for informative missing data. To be published.

Glynn, J. J., Laird, N. M. and Rubin, D. B. (1986) Selection modelling *versus* mixture modelling with nonignorable nonresponse. In *Drawing Inferences from Self Selected Samples* (ed. H. Wainer). New York: Springer.

——(1993a) The performance of mixture models for nonignorable nonresponse with follow ups. *J. Am. Statist. Ass.*, **88**, 984–993.

———(1993b) Nonignorable nonresponse with follow-ups: selection and mixture-modelling compared. *J. Am. Statist. Ass.*, **88**, 984–993.
Gould, A. L. (1980) A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics*, **36**, 721–727.
Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982) Imputation of missing values with the probability of response depends on the variable being imputed. *J. Am. Statist. Ass.*, **77**, 251–261.
Gruger, J., Kay, R. and Schumacher, M. (1991) The validity of inferences based on incomplete observations in disease state models. *Biometrics*, **47**, 595–605.
Hausman, J. A. and Wise, D. A. (1979) Attrition bias in experimental and panel data: the Gary Income Maintenance Experiment. *Econometrica*, **46**, 455–473.
Heckman, J. (1976) The common structure of statistical models of truncation, sample selection, limited dependent variables and a simple estimator for such models. *Ann. Econ. Socl Measmnt*, **5**, 475–492.
Heitjan, D. F. (1993) Ignorability and coarse data: some biomedical examples. *Biometrics*, to be published.
Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *Ann. Statist.*, **19**, 2244–2253.
Henderson, R. and Oman, P. (1993) Influence in linear hazard models. *Scand. J. Statist.*, **20**, in the press.
Heyting, A., Essers, J. G. A. and Tolboom, J. T. B. M. (1990) A practical application of the Patel–Kenward analysis of covariance to data from an anti-depressant trial with drop-outs. *Statist. Appl.*, **2**, 295–307.
Heyting, A. and Tolboom, J. T. B. M. (1993) Authors' reply. *Statist. Med.*, to be published.
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.
Hsiao, C. (1986) *Analysis of Panel Data*, sect. 8.3. Cambridge: Cambridge University Press.
Jewell, N. P. and Nielson, J. P. (1993) A framework for consistent prediction rules based on markers. *Biometrika*, **80**, 153–164.
Kay, R. (1986) A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics*, **42**, 855–865.
Laird, N. M. (1988) Missing data in longitudinal studies. *Statist. Med.*, **7**, 305–315.
Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
Lavori, P. W. (1992) Clinical trials in psychiatry: should protocol deviations censor patient data (with discussion)? *Neuropsychopharmacology*, **6**, 39–63.
Little, R. J. A. (1982) Models for nonresponse in sample surveys. *J. Am. Statist. Ass.*, **77**, 237–250.
———(1983) Superpopulation models for nonresponse. In *Incomplete Data in Sample Surveys* (eds W. G. Madow, I. Olkin and D. B. Rubin), vol. 2. New York: Academic Press.
———(1985) A note about models for selectivity bias. *Econometrica*, **53**, 1469–1474.
———(1993a) Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Ass.*, **88**, 125–134.
———(1993b) A class of pattern-mixture models for normal missing data. *Biometrika*, to be published.
Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
Little, R. J. A. and Wang, Y. (1993) Pattern-mixture models for multivariate incomplete data with covariates. *Joint Statistical Meetings, San Francisco*.
Lombard, M., Portmann, B., Neuberger, J., Williams, R., Tygstrup, N., Ranek, L., Ring-Larsen, H., Rodes, J., Navasa, M., Trepo, C., Pape, G., Schou, G., Badsberg, J. H. and Andersen, P. K. (1993) Cyclosporin A treatment in primary biliary cirrhosis: results of a long-term placebo controlled trial. *Gastroenterology*, **104**, 519–526.
Mardia, K. V. and Goodall, C. (1993) Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics* (eds G. P. Patil and C. R. Rao). Amsterdam: North-Holland.
Mardia, K. V. and Little, J. (1993) Thin plate splines and kriging with derivatives information. To be published.
Mori, M., Woolson, R. F. and Woodworth, G. G. (1991) Slope estimation in the presence of informative right censoring: modeling the number of observations as a geometric random variable. *Biometrics*, to be published.
Nelder, J. A. and Mead, R. (1965) A simplex method for function minimisation. *Comput. J.*, **7**, 303–313.
Okafor, R. (1982) Bias due to logistic regression in sample surveys. *PhD Thesis*. Harvard University, Cambridge.
O'Neill, R. (1971) Algorithm AS 47: Function minimization using a simplex procedure. *Appl. Statist.*, **20**, 338–345.
Patterson, H. D. and Thompson, R. (1971) Recovery of interblock information when block sizes are unequal. *Biometrika*, **54**, 545–554.
Phillips, G. R. and Eyring, E. M. (1988) Error estimation using the sequential simplex method in nonlinear least squares data analysis. *Anal. Chem.*, **60**, 738–741.
Prentice, R. L. (1990) Opportunities for enhancing efficiency and reducing cost in large scale disease prevention trials: a statistical perspective. *Statist. Med.*, **9**, 161–172.
Richardson, S. and Gilks, W. R. (1993) Conditional independence models for epidemiological studies with covariate measurement error. *Statist. Med.*, **12**, 1703–1722.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
———(1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
———(1992) Computational aspects of analyzing random effects/longitudinal models. *Statist. Med.*, **11**, 1809–1821.
Rubin, D. B. and Schenker, N. (1991) Multiple imputation in health-care data bases: an overview and some applications. *Statist. Med.*, **10**, 585–598.
Schluchter, M. D. (1992) Methods for the analysis of informatively censored longitudinal data. *Statist. Med.*, **11**, 1861–1870.

Shih, W. J. (1992) On informative and random dropouts in longitudinal studies. *Biometrics*, **48**, 971–972.
Shih, W. J., Quan, H. and Chang, M. N. (1993) Estimation of the mean when data contain non-ignorable missing values from a random effects model. *Statist. Probab. Lett.*, **19**, in the press.
Stefanski, L. A. and Carroll, R. A. (1985) Covariate measurement error logistic regression. *Ann. Statist.*, **13**, 1335–1351.
Tu, X. M., Meng, X. L. and Pagano, M. (1993) The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Am. Statist. Ass.*, **88**, 26–36.
Tukey, J. W. (1986) Comments on "Alternative methods for solving the problem of selection bias" by J. D. Heckman and R. Robb. In *Drawing Inferences from Self Selected Samples* (ed. H. Wainer), pp. 108–110. New York: Springer.
Verbyla, A. P. and Cullis, B. R. (1990) Modelling in repeated measures experiments. *Appl. Statist.*, **39**, 341–356.
———(1992) The analysis of multistratum and spatially correlated repeated measures data. *Biometrics*, **48**, 1015–1032.
Wu, M. C. and Bailey, K. (1988) Analyzing changes in the presence of informative right censoring caused by death and withdrawal. *Statist. Med.*, **7**, 337–346.
———(1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939–955.
Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.
Wu, M. C., Hunsberger, S. and Zucker, D. (1993) Testing for changes in the presence of censoring: parametric and nonparametric methods. *Statist. Med.*, to be published.