Appl. Statist. (2007)
56, *Part* 4, *pp.* 1–31

Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal

Peter Diggle,

Lancaster University, UK, and Johns Hopkins University School of Public Health, Baltimore, USA

Daniel Farewell

Cardiff University, UK

and Robin Henderson

University of Newcastle, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, April 18th, 2007*, Professor T. J. Sweeting *in the Chair*]

Summary. The problem of analysing longitudinal data that are complicated by possibly informative drop-out has received considerable attention in the statistical literature. Most researchers have concentrated on either methodology or application, but we begin this paper by arguing that more attention could be given to study objectives and to the relevant targets for inference. Next we summarize a variety of approaches that have been suggested for dealing with drop-out. A long-standing concern in this subject area is that all methods require untestable assumptions. We discuss circumstances in which we are willing to make such assumptions and we propose a new and computationally efficient modelling and analysis procedure for these situations. We assume a dynamic linear model for the expected increments of a constructed variable, under which subject-specific random effects follow a martingale process in the absence of drop-out. Informal diagnostic procedures to assess the tenability of the assumption are proposed. The paper is completed by simulations and a comparison of our method and several alternatives in the analysis of data from a trial into the treatment of schizophrenia, in which approximately 50% of recruited subjects dropped out before the final scheduled measurement time.

Keywords: Additive intensity model; Counterfactuals; Joint modelling; Martingales; Missing data

1. Introduction

Our concern in this paper is with longitudinal studies in which a real-valued response Y is to be measured at a prespecified set of time points, and the target for inference is some version of the expectation of Y. Studies of this kind will typically include covariates X, which may be time constant or time varying. Frequently, the interpretation of the data is complicated by *drop-outs*: subjects who are lost to follow-up before completion of their intended sequence of measurements. The literature on the analysis of longitudinal data with drop-outs is extensive: important early references include Laird (1988), Wu and Carroll (1988) and Little (1995), for which the *Web of Science* lists approximately 200, 170 and 300 citations respectively, up to the end of 2006.

Address for correspondence: Daniel Farewell, Department of Epidemiology, Statistics and Public Health, Centre for Health Sciences Research, Cardiff University, Neuadd Meirionnydd, Heath Park, Cardiff, CF14 4YS, UK.
E-mail: farewelld@cf.ac.uk

 0035–9254/07/56000

A useful classification of drop-out mechanisms is the hierarchy that was introduced by Rubin (1976) in the wider context of missing data. Drop-out is *missing completely at random* (MCAR) if the probability that a subject drops out at any stage depends neither on their observed responses nor on the responses that would have been observed if they had not dropped out. Drop-out is *missing at random* (MAR) if the probability of drop-out may depend on observed responses but, given the observed responses, is conditionally independent of unobserved responses. Drop-out is *missing not at random* (MNAR) if it is not MAR. Note that we interpret drop-out MCAR, MAR and MNAR only as properties of the joint distribution of random variables representing a sequence of responses Y and drop-out indicators R; Little (1995) developed a finer classification by considering also whether drop-out does or does not depend on covariates X. From the point of view of inference, the importance of Rubin's classification is that, in a specific sense that we discuss later in the paper, likelihood-based inference for Y is valid under drop-out MAR, whereas other methods for inference, such as the original form of generalized estimating equations (Liang and Zeger, 1986), require drop-out MCAR for their validity. Note also that, if the distributional models for the responses Y and drop-out indicators R include parameters in common, likelihood-based inference under drop-out MAR is potentially inefficient; for this reason, the combination of drop-out MAR and separate parameterization is sometimes called *ignorable*, and either drop-out MNAR or MAR with parameters in common is sometimes called *non-ignorable* or *informative*. The potential for confusion through different interpretations of these terms is discussed in a chain of correspondence by Ridout (1991), Shih (1992), Diggle (1993) and Heitjan (1994).

Our reasons for revisiting this topic are threefold. Firstly, we argue that in the presence of drop-outs the inferential objective is often defined only vaguely. Though there are other possibilities, the most common target is the mean response, which we also adopt. However, many possible expectations are associated with Y: in Section 2 we contend that, in different applications, the target may be one of several unconditional or conditional expectations. However, in *all* applications careful thought needs to be given to the purpose of the study and the analysis, with recognition that drop-out leads to missing data but should not be considered solely as an *indicator* of missingness. The common notation $Y = (Y_{\text{obs}}, Y_{\text{miss}})$ blurs this distinction. The complexity of some of the models and methods that are now available in the statistics literature may obscure the focus of a study and its precise objective under drop-out. For this reason, we use as a vehicle for discussion the very simple setting of a longitudinal study with only two potential follow-up times and one drop-out mechanism. A second but connected issue is that the assumptions underlying some widely used methods of analysis are subtle; Section 3 provides a discussion of these assumptions and an overview of the development of some of the important methodology. We discuss what can and cannot be achieved in practice, again by using the two-time-point scenario for clarity. Our third purpose in this paper is to offer in Section 4 an approach that is based on dynamic linear models for the expected increments of the longitudinal process. The assumptions on which we base our models are easily stated and doubly weak: weak with respect to both longitudinal and drop-out processes. None-the-less, all methods for dealing with missing data require, to some extent, untestable assumptions, and ours is no exception. However, we are willing to make such assumptions in the following circumstances. Firstly, the targets for inference are parameters of a hypothetical drop-out-free world that describes what would have happened if the drop-out subjects had in fact continued. Secondly, any unexplained variability between subjects exhibits a certain stability before drop-out. Thirdly, such stability is maintained beyond each drop-out time by the diminishing subset of continuing subjects.

The first point is discussed in Section 2 and the 'stability' requirement of the next two points is defined formally in Section 4 as a martingale random-effects structure. Section 4 also presents

graphical diagnostics and an informal test procedure for critical assessment of this property. Our methods are quite general but for discussion purposes we return to the two-time-point scenario in Section 5, before demonstrating the methods through simulations in Section 6. Section 7 describes a comparative analysis of data from a trial into the treatment of schizophrenia. The paper closes with brief discussion in Section 8. Appendix A describes an implementation of our proposal in the S language.

Our topic can be regarded as a special case of a wider class of problems concerning the joint modelling of a longitudinal sequence of measured responses and times to events. Longitudinal data with drop-out can formally be considered as joint modelling in which the time to event is the drop-out time as, for example, in Henderson *et al.* (2000). In Section 7, we reanalyse the data from their clinical example to emphasize this commonality and to illustrate our new approach. For recent reviews of joint modelling, see Hogan *et al.* (2004) or Tsiatis and Davidian (2004).

Under our new approach, estimators are available in closed form and are easily interpretable. Further, estimation is computationally undemanding, as processing essentially involves a least squares fit of a linear model at each observation time. This is in contrast with many existing approaches to drop-out prone data where, in our experience, the computational load of model fitting can be a genuine obstacle to practical implementation when the data have a complex structure and there is a need to explore a variety of candidate models.

2. Inferential objectives in the presence of drop-out

As indicated in Section 1, we consider in this section a study involving a quantitative response variable Y, which can potentially be measured at two time points $t = 1, 2$ but will not be measured at $t = 2$ for subjects who drop out of the study. We ignore covariate effects and focus on estimation of $\mu_t = \mathbf{E}(Y_t)$, though similar arguments apply to the full distributions of the response variables. We emphasize that this simple setting is used only to illustrate underlying concepts without unnecessary notational complication. The general thrust of the argument applies equally to more elaborate settings.

At time 1 the response is observed for all subjects, but at time 2 the response may be missing owing to drop-out. Leaving aside for the moment the scientific purpose of the study and concentrating on statistical aspects, it is tempting to begin with the model

$$\left.\begin{aligned} Y_1 &= \mu_1 + Z_1, \\ Y_2 &= \mu_2 + Z_2, \\ \mathbf{E}(Z_1) &= \mathbf{E}(Z_2) = 0. \end{aligned}\right\} \tag{1}$$

The parameter μ_1 is the population mean at time 1. Writing down model (1) invites a similar interpretation for μ_2. In fact, the apparently straightforward adoption of model (1) brings with it some interesting but usually unstated or ignored issues.

For the moment we ignore context and consider four abstract random variables, which we shall call Y_1, Y_{2a}, Y_{2b} and R, the last of which is binary. Our primary interest is in the expectations of the Y-variables, and we write

$$\left.\begin{aligned} & & Y_{2a} &= \mu_{2a} + Z_{2a}, & \\ Y_1 &= \mu_1 + Z_1, & \mathbf{P}(R = 0|\mathcal{S}) &= \pi(\mathcal{S}), & \\ & & Y_{2b} &= \mu_{2b} + Z_{2b}. & \end{aligned}\right\} \tag{2}$$

In expression (2), $\mathbf{E}(Z_1) = \mathbf{E}(Z_{2a}) = \mathbf{E}(Z_{2b}) = 0$, $\mathcal{S}$ denotes a set of conditioning variables and we allow $\pi(\cdot)$ to depend arbitrarily on $\mathcal{S}$. We make no assumption of independence between

Z_1, Z_{2a} and Z_{2b}, and for the unconditional case $\mathcal{S} = \emptyset$ we write $\pi = \pi(\emptyset) = \mathbf{P}(R = 0)$. By construction, the parameters μ_1, μ_{2a} and μ_{2b} are the marginal expectations of Y_1, Y_{2a} and Y_{2b} respectively.

In the context of longitudinal data with drop-outs, subjects with $R = 1$ are the completers, who are denoted group $\mathcal{C}$. For each completer, Y_1, Y_{2a} and R are observed and have the obvious interpretations as the responses at times 1 and 2 together with an indicator of response, whereas Y_{2b} is an unobserved counterfactual, representing the value of the response that would have been observed if the subject had in fact dropped out.

The drop-outs, group $\mathcal{D}$, are those subjects who have $R = 0$. These subjects experience the event of dropping out of the study, which in different contexts may mean discontinuation of treatment, cessation of measurement or both. If drop-out refers only to the discontinuation of treatment, then Y_{2b} is the observed response at time 2, and Y_{2a} the counterfactual that would have been observed if the subject had continued treatment. This situation, where drop-out does not lead to cessation of measurement, is one which we discuss no further. Throughout the remainder of the paper, we are concerned with the case when $R = 0$ does correspond to cessation of measurement, and consequently neither Y_{2a} nor Y_{2b} is observed for any subject in group $\mathcal{D}$. In this case, Y_{2b} is the extant, but unobserved, longitudinal response at time 2 and Y_{2a} is the counterfactual that would have been observed if the subject in question had not dropped out.

In this framework we make explicit the possibility that the act of dropping out can influence the response, rather than simply lead to data being missing. In other words, we separate the consequence of dropping out from the observation of that consequence. At least conceptually, the events 'avoiding drop-out' and 'observing Y_{2a}' are considered to be distinct.

The above is reminiscent of the usual framework for causal inference, as described for instance by Rubin (1991) or Rubin (2004), in which R would be a binary treatment assignment or other intervention indicator. However, there are three important differences. The most obvious is that with drop-out we *never* observe Y_{2b}, whereas in causal inference it would be observed for each subject in group $\mathcal{D}$. The second difference is that, assuming no initial selection effect, in the longitudinal setting we observe Y_1 for all subjects, and this can be exploited in inference through assumed or estimated relationships between responses before and after drop-out. The third difference is that we assume R to be intrinsic to the subject rather than an assigned quantity such as treatment, and between-subject independence is sufficient for us to avoid the need to discuss assignment mechanisms.

In particular applications we need to consider the scientific objective of the study and consequent target for inference. At time $t = 1$ we can easily estimate $\mu_1 = \mathbf{E}(Y_1)$ by standard techniques. Our focus will be the target for estimation at time $t = 2$, which we assume can be expressed as some property of a random variable Y_2, typically $\mathbf{E}(Y_2)$. We discuss this within the specific setting of model (2).

2.1. Objective 1: realized second response

The first possible target for inference that we discuss is the realized, non-counterfactual, second response

$$Y_2 := Y_{2a}R + Y_{2b}(1 - R), \tag{3}$$

which is unobserved for subjects in group $\mathcal{D}$. Further progress will therefore depend on the strong and untestable assumption that $Y_{2a} = Y_{2b}$. This assumption seems to be implicit in most published work and may be reasonable in circumstances where drop-out is deemed to have no material effect on the measurement other than causing it to be missing. Applied uncritically,

however, this can result in misleading inference about Y_2. For example, drop-out might be because of death, in which case Y_{2b} could be assigned an arbitrary value such as 0 and the definition of Y_2 above is, for practical purposes, meaningless.

In contrast, the data that we analyse in Section 7 come from a longitudinal randomized clinical trial of drug treatments for schizophrenia, in which drop-out implies discontinuation of the assigned drug and the response could have been (but in fact was not) measured after drop-out. In this setting, Y_2 as defined at expression (3), is readily interpretable as the intention-to-treat response.

2.2. Objective 2: conditional second response

A second possible target for inference is the response at time $t = 2$ conditional on not dropping out, or equivalently

$$Y_2 := \begin{cases} Y_{2a} & \text{if } R = 1, \\ \text{undefined} & \text{if } R = 0. \end{cases}$$

Only complete cases, group $\mathcal{C}$, contribute to inference, which is therefore always conditional on $R = 1$. This is perfectly proper if the objective is to study the response within the subpopulation of subjects who do not drop out.

In the schizophrenia example, some subjects were removed from the study because their condition did not improve. Objective 2 would therefore be appropriate in this context if interest were confined to the subset of subjects who had not yet been removed from the study owing to inadequate response to treatment.

2.3. Objective 3: hypothetical second response

Our third potential target for inference, again unobserved for group $\mathcal{D}$ subjects, is

$$Y_2 := Y_{2a},$$

which is appropriate if scientific interest lies in the (possibly hypothetical) response distribution of a drop-out-free population. We note that this is analogous to the usual estimand in event history analysis, with drop-out equivalent to censoring. The assumption $Y_{2a} = Y_{2b}$ makes objectives 1 and 3 equivalent.

The essential difference between the interpretations of Y_2 under objectives 2 and 3 is between the *marginal* and *conditional* distributions of the response at time 2. This can be substantial, as would be the case if, for example, drop-out occurs if and only if $Z_{2a} < 0$. This might seem an extreme example, but it could never be identified from the observed data.

It is important that the objectives be clearly stated and understood at the outset of a study, especially for regulatory purposes. There are similarities with distinguishing intention-to-treat and per-protocol analyses (Sommer and Zeger, 1991; Angrist *et al.*, 1996; Little and Yau, 1996; Frangakis and Rubin, 1999) and with causal inference in the presence of missing data or non-compliance quite generally (Robins, 1998; Peng *et al.*, 2004; Robins and Rotnitzky, 2004). The hypothetical second response Y_{2a} will be our inferential target for the analysis that we present in Section 7 for the schizophrenia data. We argue that in this setting, where drop-out need not be related to an adverse event, clinical interest genuinely lies in the hypothetical response that patients would have produced *if they had not dropped out*. This is likely to be of greater value than the realized or conditional second responses, since treatment performance is of more concern than subject profiles. We emphasize, however, that this need not always be the so, and that in some circumstances a combination of objectives may be appropriate. For example, Dufoil *et al.*

(2004) and Kurland and Heagerty (2005) separately discussed applications in which there are two causes of drop-out: death and possibly informative loss to follow-up. In these applications the appropriate target for inference is the response distribution in the *hypothetical* absence of loss to follow-up but *conditional* on not dying, thus combining objectives 2 and 3. In other applications it is quite possible that a combination of all three objectives may be appropriate.

3. Approaches to the analysis of longitudinal data with drop-out

We now illustrate in the context of model (2) some of the variety of approaches that have been proposed for the analysis of longitudinal data with drop-out. We do not attempt a complete review (see Hogan and Laird (1997a, b), Little (1998), Hogan *et al.* (2004), Tsiatis and Davidian (2004) or Davidian *et al.* (2005)) but hope to give a flavour of the broad classes of methods and their underlying assumptions.

3.1. Complete case

Complete-case analysis is probably the simplest approach to dealing with drop-outs, as we simply ignore all non-completers. As discussed earlier, this is appropriate for objective 2, or in more formal language when our interest lies in the conditional distribution $[Y_1, Y_{2a}|R=1]$. The relevant estimator within model (2) is

$$\bar{Y}_{2a}^{\mathcal{C}} = \frac{1}{|\mathcal{C}|}\sum_{\mathcal{C}} Y_{2a},$$

which estimates

$$\mu_{2a} + \mathbf{E}(Z_{2a}|R=1).$$

3.2. Pattern–mixture

A complete-case analysis forms one component of a pattern–mixture approach (Little, 1993), in which we formulate a separate submodel for each of $[Y_1|R=0]$ and $[Y_1, Y_{2a}|R=1]$, perhaps with shared parameters. From this, we can obtain valid inference for the marginal $[Y_1]$ by averaging, but again only conditional inference for $[Y_{2a}|R=1]$, as with complete-case analysis. The pattern–mixture approach is intuitively appealing from the perspective of retrospective data analysis, in which context it is natural to compare response distributions in subgroups that are defined by different drop-out times. From a modelling perspective it is also natural if we regard the distribution of R as being determined by latent characteristics of the individual subjects. In its most general form, the pattern–mixture approach is less natural if we regard drop-out as a consequence of a subject's response history, because it allows conditioning on the future. However, Kenward *et al.* (2003) discussed the construction of pattern–mixture specifications that avoid dependence on future responses.

3.3. Imputation methods

Imputation methods implictly focus on objective 3, sometimes adding the assumption that $Y_{2a} = Y_{2b}$, in which case objectives 1 and 3 are equivalent.

3.3.1. Last observation carried forward

The last observation carried forward (LOCF) method imputes Y_{2a} by Y_1 for each subject in group $\mathcal{D}$. Writing $\hat{\pi} = |\mathcal{D}|/n$, the implied estimator for the mean response at time 2 is $\bar{Y}_{2a}^{\mathcal{C}}(1-\hat{\pi}) + \bar{Y}_1^{\mathcal{D}}\hat{\pi}$, where $\bar{Y}_1^{\mathcal{D}}$ is the mean at time 1 for group $\mathcal{D}$. The estimator is consistent for

$$\mu_{2a}(1-\pi)+\mu_1\pi+\mathbf{E}[Z_{2a}\{1-\pi(Z_{2a})\}]+\mathbf{E}\{Z_1\pi(Z_1)\}$$

and hence is not obviously useful. The LOCF method is temptingly simple and is widely used in pharmaceutical trials, but it has attracted justifiable criticism (Molenberghs *et al.*, 2004).

3.3.2. Last residual carried forward

A variant of the LOCF method would be to carry forward a suitably defined residual. Suppose, for example, that we define

$$Y_2=\begin{cases} Y_{2a} & \text{if } R=1,\\ \bar{Y}_{2a}^{\mathcal{C}}+(Y_1-\bar{Y}_1) & \text{if } R=0.\end{cases}$$

The implicit estimator is then

$$\bar{Y}_2=\bar{Y}_{2a}^{\mathcal{C}}-(1-\hat{\pi})(\bar{Y}_1^{\mathcal{C}}-\bar{Y}_1), \tag{4}$$

which is consistent for $\mu_{2a}+\mathbf{E}(Z_{2a}|R=1)-(1-\pi)\,\mathbf{E}(Z_1|R=1)$. Typically, if completers were high responders at time 1, then we might expect the same to apply at time 2, and vice versa. The variables Z_1 and Z_{2a} would then have the same sign. The expectation of $\bar{Y}_2$ will be closer to μ_{2a} than the expectation of $\bar{Y}_{2a}^{\mathcal{C}}$, which is a desirable shift from the complete-case estimand if μ_{2a} is the target for inference.

For these reasons the last residual carried forward method must be preferable to the LOCF approach as a means of overcoming potentially informative drop-out, but in our opinion it does not provide an adequate solution to the problem. We describe it here principally to highlight two important points. Firstly, the unspoken question underlying the estimator (4) is 'how unusual were the completers at time 1?'. If they were unusual, then we presume that this may also have been true at time 2, and consequently adjust the observed time 2 average accordingly. Second, this adjustment is downweighted by a factor $1-\hat{\pi}$. We observe, anticipating results in Section 4, that in our hypothetical drop-out-free universe $\pi=0$, suggesting the estimator $\bar{Y}_{2a}^{\mathcal{C}}-(\bar{Y}_1^{\mathcal{C}}-\bar{Y}_1)$ as another alternative.

3.3.3. Multiple imputation

One of several possible criticisms of both the LOCF and the last residual carried forward methods is that, at best, they ignore random variation by imputing fixed values. Hot deck imputation addresses this by sampling post-drop-out values from a distribution; in principle, this could be done either by sampling from an empirical distribution, such as that of the observed values from other subjects who did not drop out but had similar values of available explanatory variables, or by simulating from a distributional model. Multiple-imputation methods (Rubin, 1987) take this process one step further, by replicating the imputation procedure to enable estimation of, and if necessary adjustment for, the component of variation that is induced by the imputation procedure.

3.4. Missing at random: parametric modelling

Any assumed parametric form for the joint distribution $[Y_1, Y_{2a}, R]$ cannot be validated empirically, because we can check only the marginal $[Y_1]$ and conditional $[Y_1, Y_{2a}|R=1]$ distributions. The assumption of drop-out MAR is useful because it allows one part of the joint distribution to remain unspecified. This assumption assumes that the probability of drop-out does not depend on the outcome at time 2 given the value at time 1, whence $\pi(Y_1, Y_{2a}, Y_{2b})$ simplifies to $\pi(Y_1)$. In general this assumption is untestable, but if we combine it with a parametric model for $[Y_1, Y_{2a}]$

we obtain the beguiling result that likelihood inference is possible without any need to model $\pi(Y_1)$. The likelihood contribution in the $\mathcal{C}$ group is

$$\mathbf{P}(R=1|Y_1, Y_{2a})[Y_1, Y_{2a}] = \{1-\pi(Y_1)\}[Y_1, Y_{2a}],$$

whereas in the $\mathcal{D}$ group it is just $\pi(Y_1)[Y_1]$. The combined likelihood is thus $L = L_{R|Y} L_Y$, where

$$L_{R|Y} = \prod [R|Y_1],$$
$$L_Y = \prod_{\mathcal{C}} [Y_1, Y_{2a}] \prod_{\mathcal{D}} [Y_1].$$

The factorization $[Y, R] = [R|Y][Y]$ is usually called a selection model (e.g. Michiels *et al.* (1999)), although we prefer the term *selection factorization*, to contrast with the *pattern–mixture factorization* $[Y, R] = [Y|R][R]$, and to emphasize the distinction between how we choose to model the data and how we subsequently conduct data analysis.

As an illustration, suppose that $(Z_1, Z_{2a})'$ is distributed as $N(0, \sigma^2 V)$, with

$$V = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{5}$$

Then the maximum likelihood estimator of μ_{2a} under drop-out MAR is

$$\hat{\mu}_{2a} = \bar{Y}_{2a}^{\mathcal{C}} - \hat{\rho}(\bar{Y}_1^{\mathcal{C}} - \bar{Y}_1), \tag{6}$$

which again adjusts the observed time 2 sample mean according to how unusual the fully observed group were at time 1, with shrinkage. Once more we call attention to this estimator, and note an interpretation of the estimator $\bar{Y}_{2a}^{\mathcal{C}} - (\bar{Y}_1^{\mathcal{C}} - \bar{Y}_1)$ as being appropriate when within-subject variability is small ($\rho \to 1$).

Parametric modelling under the combined assumption of drop-out MAR and separate parameterization has the obvious attraction that a potentially awkward problem can be ignored and likelihood-based inference using standard software is straightforward. A practical concern with this approach is that the ignorability assumption is untestable without additional assumptions. A more philosophical concern arises if, as is usually so, the data derive from discrete time observation of an underlying continuous time process. In these circumstances, it is difficult to imagine any mechanism, other than adminstrative censoring, under which drop-out at time t could depend on the observed response at time $t-1$ but not on the unobserved response trajectory between $t-1$ and t.

3.5. Missing at random: unbiased estimating equations

If interest is confined to estimating μ_{2a}, or more generally covariate effects on the mean, then an alternative approach, which is still within the framework of drop-out MAR, is to model $\pi(Y_1)$ but to leave $[Y_1, Y_{2a}]$ unspecified.

Under drop-out MAR we can estimate the probability of drop-out consistently from the observed data: we need only R and Y_1 for each subject, both of which are always available. This leads to an estimated $\hat{\pi}(Y_1)$ of drop-out probability, often via a logistic model. The marginal mean of Y_{2a} can now be estimated consistently by using a weighted average of the observed Y_{2a}, where the weights are the inverse probabilities of observation (Horvitz and Thompson, 1952; Robins *et al.*, 1995):

$$\hat{\mu}_{2a} = \sum_{\mathcal{C}} \frac{Y_{2a}}{1-\hat{\pi}(Y_1)} \Big/ \sum_{\mathcal{C}} \frac{1}{1-\hat{\pi}(Y_1)}. \tag{7}$$

Use of equation (7) requires $1-\hat{\pi}(Y_1)$ to be strictly positive for all subjects, and it encounters difficulties in practice if this probability can be close to 0. This will not often be a material restriction within the current simplified setting, but it can be problematic in more complex study designs with high probabilities of drop-out in some subgroups of subjects.

3.6. Missing not at random: Diggle–Kenward model

Diggle and Kenward (1994) discussed a parametric approach to the problem of analysing longitudinal data with drop-outs, based on a selection factorization. In the special case of model (2), the Diggle and Kenward model reduces to $(Z_1, Z_2)' \sim N(0, \sigma^2 V)$ with V as in equation (5), and

$$\pi(Y_1, Y_2) = \frac{\exp(\alpha + \gamma_1 Y_1 + \gamma_0 Y_2)}{1+\exp(\alpha + \gamma_1 Y_1 + \gamma_0 Y_2)}, \tag{8}$$

with the tacit assumption that $Y_2 = Y_{2a} = Y_{2b}$. Drop-out is MAR if $\gamma_0 = 0$ and MCAR if $\gamma_0 = \gamma_1 = 0$. The model therefore maps directly onto Rubin's hierarchy, and in particular MAR drop-out is a parametrically testable special case of a drop-out MNAR model. Although the likelihood does not separate in the same way as under parametric drop-out MAR, likelihood inference is still possible by replacing π with its conditional expectation, which is derived from the conditional distribution of Y_2 given Y_1. The price that is paid for this facility is that correct inference now depends on two untestable modelling assumptions, the normal distribution model for (Y_1, Y_2) and the logistic model for drop-out (Kenward, 1998). There is no closed form for the estimator of μ_{2a}.

3.7. Missing not at random: random effects

Under the Diggle and Kenward model the probability of drop-out is directly determined by the responses Y_1 and Y_2, again assuming that $Y_{2a} = Y_{2b}$. If measurement error contributes substantially to the distribution of Y, a random-effects model may be more appealing. In this approach, the usual modelling assumption is that Y and R are conditionally independent given shared, or more generally dependent, random effects. See, for example, Wu and Carroll (1988), Little (1995), Berzuini and Larizza (1996), Wulfsohn and Tsiatis (1997), Henderson *et al.* (2000) and Xu and Zeger (2001). A simple model for our simple example is

$$\begin{aligned} Y_1 &= \mu_1 + U + \varepsilon_1, \\ Y_2 &= \mu_2 + U + \varepsilon_2, \\ U &\sim N(0, \tau^2), \\ \varepsilon_1, \varepsilon_2 &\sim N(0, \sigma^2), \\ \pi(U, \varepsilon_1, \varepsilon_2) = \pi(U) &= \frac{\exp(\alpha + \gamma U)}{1+\exp(\alpha + \gamma U)} \end{aligned}$$

with independence between U, ε_1 and ε_2. Models of this type are in general drop-out MNAR models, because random effects are always unobserved and typically influence the distribution of Y at all time points. It follows that the conditional distribution of the random effects, and hence the probability of drop-out given Y, depends on the values of Y at all time points, and in particular on values that would have been observed if the subject had not dropped out.

For maximum likelihood estimation for the simple model above, the shared effect U can be treated as missing data and methods such as the EM or Markov chain Monte Carlo algorithms used, or the marginal likelihood can be obtained by numerical integration over U, and the resulting likelihood maximized directly. Implementation is computationally intensive, even for this simple example, and there is again no closed form for $\hat{\mu}_{2a}$.

Models of this kind are conceptually attractive, and parameters are identifiable without any further assumptions. But, as with the Diggle–Kenward model, the associated inferences rely on distributional assumptions which are generally untestable. Furthermore, in our experience the computational demands can try the patience of the statistician.

3.8. Missing not at random: unbiased estimating equations

A random-effects approach to joint modelling brings yet more untestable assumptions and we can never be sure that our model is correct for the unobserved data, although careful diagnostics can rule out models that do not even fit the observed data (Dobson and Henderson, 2003). Rotnitzky *et al.* (1998), in a follow-up to Robins *et al.* (1995), argued strongly for a more robust approach, on the assumption that the targets for inference involve only mean parameters. They again left the joint distribution of responses unspecified but now modelled the drop-out probability as a function of both Y_1 and Y_{2a}, e.g. by the logistic model (8). As applied within the simple framework of model (2), the most straightforward version of the procedure of Rotnitzky *et al.* (1998) is two stage: first, estimate the drop-out parameters from an unbiased estimating equation; second, plug drop-out probability estimates into another estimating equation.

For example, the drop-out parameters α, γ_0 and γ_1 in equation (8) might be estimated by solving

$$\sum_{\mathcal{C}} \frac{\hat{\pi}(Y_1, Y_{2a})}{1-\hat{\pi}(Y_1, Y_{2a})}\,\phi(Y_1) - \sum_{\mathcal{D}} \phi(Y_1) = 0, \tag{9}$$

where $\phi(Y_1)$ is a user-defined vector-valued function of Y_1. As there are three unknowns in our example, $\phi(Y_1)$ needs to be three dimensional, such as $\phi(Y_1) = (1, Y_1, Y_1^2)'$. Since we need only $\pi(Y_1, Y_{2a})$ in the fully observed group, all components of equation (9) are available, and for estimation there is no need for assumptions about Y_{2b}. Assumptions would, however, be needed for estimands to be interpretable. Rewriting equation (9) as

$$\sum \left\{ \mathbf{1}(R=1) \frac{\hat{\pi}(Y_1, Y_{2a})}{1-\hat{\pi}(Y_1, Y_{2a})} - \mathbf{1}(R=0) \right\} \phi(Y_1) = 0,$$

it is easy to see that the equation is unbiased by taking conditional expectations of the indicator functions given (Y_1, Y_{2a}).

At the second stage, the newly obtained estimated drop-out probabilities are plugged into an inverse probability weighted estimating equation to give

$$\hat{\mu}_{2a} = \sum_{\mathcal{C}} \frac{Y_{2a}}{1-\hat{\pi}(Y_1, Y_{2a})} \Big/ \sum_{\mathcal{C}} \frac{1}{1-\hat{\pi}(Y_1, Y_{2a})}.$$

Rotnitzky *et al.* (1998) indicated that efficiency can be improved by augmenting the estimating equation for μ_{2a} by a version of equation (9) (with a different ϕ) and simultaneously solving both equations for all parameters. Fixed weight functions may also be introduced as usual. They also argued that estimation of the informative drop-out parameter γ_0 will be at best difficult and that the validity of the drop-out model cannot be checked if $\gamma_0 \neq 0$. Their suggestion is that

γ_0 be treated as a known constant but then varied over a range of plausible values to assess sensitivity of inferences for other parameters to the assumed value of γ_0.

Carpenter *et al.* (2006) compared inverse probability weighting (IPW) methods with multiple imputation. In particular, they considered a doubly robust version of IPW, which was introduced by Scharfstein *et al.* (1999) in their rejoinder to the discussion, which gives consistent estimation for the marginal mean of Y_{2a} provided that at most one of the models for R or for Y_{2a} is misspecified. Their results show that doubly robust IPW outperforms the simpler version of IPW when the model for R is misspecified, and it outperforms multiple imputation when the model for Y_{2a} is misspecified.

3.9. Sensitivity analysis

Rotnitzky *et al.* (1998) are not the only researchers to suggest sensitivity analysis in this context. Other contributions include Copas and Li (1997), Scharfstein *et al.* (1999, 2003), Kenward (1998), Rotnitzky *et al.* (2001), Verbeke *et al.* (2001), Troxel *et al.* (2004), Copas and Eguchi (2005) and Ma *et al.* (2005).

Sensitivity analysis with respect to a parameter that is difficult to estimate is clearly a sensible strategy and works best when the sensitivity parameter is readily interpretable in the sense that a subject-matter expert can set bounds on its reasonable range; see, for example, Scharfstein *et al.* (2003). In that case, if the substantively important inferences show no essential change within the reasonable range, all is well. Otherwise, there is some residual ambiguity of interpretation.

Most parametric approaches can also be implemented within a Bayesian paradigm. An alternative to a sensitivity analysis is then a Bayesian analysis with a suitably informative prior for γ_0.

3.10. Conclusions

Existing approaches to the analysis of longitudinal data subject to drop-out may, if only implicitly, be addressing different scientific or inferential objectives. In part this may be because methods and terminology that are designed for general multivariate problems with missing data do not explicitly acknowledge the evolution over time of longitudinal data. In the next section we offer an alternative, which we believe is better suited to the longitudinal set-up and which borrows heavily from event history methodology. We consider processes evolving in time and propose a martingale random-effects model for the longitudinal responses, combined with a drop-out mechanism that is allowed to depend on both observed and unobserved history, but not on the future. The martingale assumption formalizes the idea that adjusting for missing data is a defensible strategy provided that subjects' longitudinal response trajectories exhibit stability over time. Our drop-out model is formally equivalent to the independent censoring assumption that is common in event history analysis; see, for example, Andersen *et al.* (1992). We do not claim that the model proposed is universally appropriate nor suggest that it be adopted uncritically in any application. We do, however, offer some informal diagnostic procedures that can be used to assess the validity of our assumptions.

4. Proposal

4.1. Model specification

4.1.1. Longitudinal model

We suppose that τ measurements are planned on each of n independent subjects. The measurements are to be *balanced*, i.e. the intended observation times are identical for each subject, and

without loss of generality we label these times $1, \ldots, \tau$. For the time being, let us suppose that all n subjects do indeed provide τ measurements. In the notation of Section 2, Y_a is therefore observed for every subject at every observation time, and Y_b is counterfactual in every case.

We presume that covariates are also available before each of the τ observation times. These we label X_a, noting that in theory there are also counterfactual covariates X_b: the values of covariates if a subject had dropped out. We understand X_a to be an $n \times p$ matrix process, which is constant if only base-line covariates are to be used, but potentially time varying and possibly even dependent on the history of a subject or subjects. Note that we shall write $X_a(t)$ for the particular values at time t, but that by X_a without an argument we mean the entire process, and we shall follow this same convention for other processes.

At each observation time t we acknowledge that the underlying hypothetical response may be measured with mean 0 error $\varepsilon_a(t)$. We assume that this process is independent of all others and has the property that $\varepsilon_a(s)$ and $\varepsilon_a(t)$ are independent unless $s = t$. We make no further assumptions about this error process, and in particular we do not insist that its variance is constant over time.

We denote the history of the hypothetical response processes Y_a, the potentially counterfactual covariates X_a and the measurement error process ε_a, up to and including time t, by

$$\mathcal{G}_t = \{X_a(s), Y_a(s), \varepsilon_a(s) : s = 1, \ldots, t\}.$$

We are not particularly interested in how the covariates $X_a(t)$ are obtained, but for estimation we shall require that they become known at some point before time t: possibly this is at time $t-1$, or at time 0 for base-line covariates. It is useful to formalize this requirement by way of the history

$$\mathcal{G}_{t-} = \mathcal{G}_{t-1} \cup \{X_a(t)\},$$

which can be thought of as all information pertaining to X_a, Y_a and ε_a that is available strictly before time t. Since $\mathcal{G}_t$ contains information about exogenous covariates *and* measured responses, functions of either or both may be included in the matrix X_a, allowing considerable flexibility in the specification of a model.

We argue that the expected increments in Y_a are a natural choice for statistical modelling. Asking 'What happened next?' allows us to condition on available information such as the current values of covariates and responses. Later, it will also be useful to condition on the presence or absence of subjects.

For convenience, we set $X_{ai}(0) = Y_{ai}(0) = \varepsilon_{ai}(0) = 0$ for all i, adopting the notation of continuous time processes to avoid complicated subscripts. It is possible to specify a mean model for the hypothetical response vector $Y_a = (Y_{a1}, \ldots, Y_{an})'$ in terms of the discrete time *local characteristics*

$$\mathbf{E}\{\Delta Y_a(t) | \mathcal{G}_{t-}\} = \mathbf{E}\{Y_a(t) - Y_a(t-1) | \mathcal{G}_{t-}\}$$

of the process (Aalen, 1987). The local characteristics capture the extent to which the vector process Y_a is expected to change before the next observations are recorded. Local characteristics are a generalization of the intensity of a counting process. It is often possible to specify the local characteristics in terms of linear models, and in this paper we consider models of the form

$$\mathbf{E}\{\Delta Y_a(t) | \mathcal{G}_{t-}\} = X_a(t)\, \beta(t) - \varepsilon_a(t-1) \tag{10}$$

for $t = 1, \ldots, \tau$. Setting aside for one moment the issue of measurement error, we have a linear (also referred to as additive) model $X_a(t)\, \beta(t)$ for the expected increment $\mathbf{E}\{\Delta Y_a(t) | \mathcal{G}_{t-}\}$. Linear models on the increments of a process were proposed in the counting process literature by Aalen (1978), and more recently by Fosen *et al.* (2006b) for a wider class of stochastic processes. Since

a different model is specified at each time, linear models on increments can be quite general and may incorporate random intercepts, random slopes and other, more complicated, structures. We denote by β the deterministic p-vector of *regression functions* representing the effects on the local characteristics of the covariates X_a. Recall once again that β represents the hypothetical effects of covariates, assuming that drop-out does not occur. Since β is an unspecified function of time, equation (10) can be thought of as a kind of varying-coefficient model (Hastie and Tibshirani, 1993). This type of approach for longitudinal data has been taken by others: see for example Lin and Ying (2001, 2003) or Martinussen and Scheike (2000) and Martinussen and Scheike (2006), chapter 11. The crucial distinction between their work and ours is that it is the increments, not the measured responses, that are the subject of our linear model. We then accommodate measurement error by noting that, before time t, no information is available about $\varepsilon_a(t)$, so the expected change in measurement error is simply $-\varepsilon_a(t-1)$, which is known through $\mathcal{G}_{t-}$.

Incremental models correspond, on the cumulative scale, to models where the residuals form a kind of random walk, which can be thought of as additional random effects. To see this, the notion of a *transform* from the theory of discrete stochastic processes is required. Defining the cumulative regression functions $B(t)$ by $\Sigma_{s=1}^{t}\beta(s)$, with $B(0)=0$, the transform of B by X_a, denoted $X_a \cdot B$, is given by

$$\begin{aligned}(X_a \cdot B)(t) &= \sum_{s=1}^{t} X_a(s)\{B(s)-B(s-1)\} \\ &= \sum_{s=1}^{t} X_a(s)\,\beta(s)\end{aligned}$$

and forms part of the *compensator*, or predictable component, of Y_a. Note that $X_a \cdot B$ differs from the ordinary matrix product $X_a B$ and is the discrete time analogue of a stochastic integral. The transform thus captures the cumulative consequences of covariates X_a and their effects β, both of which may vary over time.

The residual process is $M_a = Y_a - X_a \cdot B - \varepsilon_a$. This process has a property that makes it a kind of random walk: it takes zero-mean steps from a current value to a future value. More formally, for $s \leqslant t$ we have that $\mathbf{E}\{M_a(t)|\mathcal{G}_s\} = M_a(s)$, and the process is thus a *martingale*. Model (10) may therefore be appropriate when, having accounted for fixed effects and measurement error, the random effects can be modelled as a martingale.

Although their conditional mean properties may seem restrictive, martingales represent, from the modeller's perspective, a wide range of processes. Neither continuity nor distributional symmetry is required of M_a, and for our purposes its variance need only be constrained to be finite. Further, the variance of the martingale increments may change over time. Serial correlation in the M_a-process induces the same in the Y_a-process, which is often a desirable property in models for longitudinal data.

The linear increments model is, on the cumulative scale, a random-effects model for Y_a of the form

$$\begin{pmatrix}\text{measured}\\ \text{response}\end{pmatrix} = \begin{pmatrix}\text{covariate}\\ \text{effects}\end{pmatrix} + \begin{pmatrix}\text{random}\\ \text{effects}\end{pmatrix} + \begin{pmatrix}\text{measurement}\\ \text{error}\end{pmatrix}.$$

The sample vector of martingale random effects is free to be, among other things, heteroscedastic, where the variance of a martingale may change over time and between subjects, and completely non-parametric, since the distribution of a martingale need not be specified by a finite dimensional parameter. We reiterate, however, that martingale residuals impose a condition

on the mean of their distribution given their past. This single condition, of unbiased estimation of the future by the past, is sufficiently strong to be easily dismissed in many application areas—though we note that this can often be overcome by suitable adjustment of the linear model. It seems to us that in many applications an underlying martingale structure seems credible, at least as a first approximation. We reiterate that the linear model may be adapted to include summaries of previous longitudinal responses if appropriate. Including dynamic covariates, e.g. summaries of the subject trajectories to date, may sometimes render the martingale hypothesis more tenable, although the interpretation of the resulting model is problematic if observed trajectories are measured with appreciable error.

We have shown that models for the hypothetical response Y_a can be defined in terms of linear models on its increments, and that such models are quite general. At no extra cost, these comprise subject-specific, martingale random effects. We do not discuss in detail the full generality of this approach; instead, we now turn to the problem of drop-out.

4.1.2. Drop-out model

Unfortunately, not all the hypothetical longitudinal responses Y_a are observed. Rather, subject i gives rise to $1 \leqslant T_i \leqslant \tau$ measurements, i.e. we observe $Y_{ai}(1), \ldots, Y_{ai}(T_i)$. Although both the hypothetical responses $Y_{ai}(T_i+1), \ldots, Y_{ai}(\tau)$ and the realized responses $Y_{bi}(T_i+1), \ldots, Y_{bi}(\tau)$ go unobserved, we restrict our assumptions to the former.

We can also consider drop-out as a dynamic process. Let R_i denote an indicator process that is associated with subject i, with $R_i(t) = 1$ if subject i is still under observation at time t, and $R_i(t) = 0$ otherwise. We let $\mathcal{R}_t$ be the history of these indicator processes up to time t. We do not distinguish between competing types of drop-out, for instance between administrative censoring, treatment failure or death, because we need not do so to make inferences regarding the hypothetical responses Y_a.

Like the covariate processes, we assume that the drop-out processes are *predictable*, in the sense that $R_i(t)$ is known strictly before time t. More formally, we shall denote by $\mathcal{R}_{t-}$ the information that is available about drop-out before time t, and assume that $R_i(t) \in \mathcal{R}_{t-}$. Although in this instance it follows that $\mathcal{R}_{t-} = \mathcal{R}_t$, it is useful to distinguish notationally between information that is available at these different points in time. We think of R_i as a process in continuous time, but in practice we are only interested in its values at discrete time points. Predictability is a sensible philosophical assumption, disallowing the possibility that drop-out can be determined by some future, unrealized, event. Note that this does not preclude the possibility that future events might depend on past drop-out.

The second important requirement that we impose on the processes R_i is that of *independent censoring*. This terminology, though standard in event history analysis, suggests more restrictions than are in fact implied. We give the formal definition and then discuss its implications for drop-out in longitudinal studies. Recall that $\mathcal{R}_{t-}$ is the history of the drop-out process before time t. Censoring (or drop-out) is said to be independent of the hypothetical response processes Y_a if, and only if,

$$\mathbf{E}\{\Delta Y_a(t) | \mathcal{G}_{t-}, \mathcal{R}_{t-}\} = \mathbf{E}\{\Delta Y_a(t) | \mathcal{G}_{t-}\}$$

(Andersen *et al.* (1992), page 139). Independent censorship says that the local characteristics of Y_a are unchanged by additional information about who has been censored already, or by knowledge of who will, or will not, be observed at the *next* point in time. Fundamentally, this assumption ensures that the observed increments remain representative of the original sample of subjects, if drop-out had not occurred. This requirement is similar in spirit to the sequential

version of drop-out MAR (Hogan *et al.* (2004), after Robins *et al.* (1995)), which states that

$$[Y_{\mathrm{a}}(t)|Y_{\mathrm{a}}(s): s<t; X_{\mathrm{a}}(s), R(s): s\leqslant t]=[Y_{\mathrm{a}}(t)|Y_{\mathrm{a}}(s): s<t; X_{\mathrm{a}}(s): s\leqslant t].$$

We emphasize that independent censoring is a weaker assumption than sequential drop-out MAR, since the former conditions on the complete past, and not just the observed past, and so allows drop-out to depend directly on latent processes. Moreover, it is a statement about conditional means, whereas the assumption of sequential drop-out MAR concerns conditional distributions.

Having laid out our assumptions concerning the drop-out process, we make a few comments on what has not been assumed. We have not specified any model, parametric or otherwise, for the drop-out process. Consequently, the drop-out process may depend on *any* aspect of the longitudinal processes, e.g. group means, subject-specific time trends or within-subject instability. The only requirement is that this dependence is not on the *future* behaviour of Y_{a}. Though often plausible, this is usually untestable.

4.1.3. Combined model

As we have already discussed, our target for inference will be the hypothetical effects of covariates supposing, contrary to fact, that subjects did not drop out of observation. More explicitly, we seek to make inference about β in the local characteristics model,

$$\mathbf{E}\{\Delta Y_{\mathrm{a}}(t)|\mathcal{G}_{t-}\}=X_{\mathrm{a}}(t)\,\beta(t)-\varepsilon_{\mathrm{a}}(t-1),$$

for the hypothetical response Y_{a}, drawing on the T_i observed covariates $X_{\mathrm{a}i}(1),\ldots,X_{\mathrm{a}i}(T_i)$ and responses $Y_{\mathrm{a}i}(1),\ldots,Y_{\mathrm{a}i}(T_i)$ for every i.

Recall that R_i is an indicator process, 1 if subject i is still under observation. We shall write

$$R(t)=\begin{pmatrix} R_1(t) & 0 & \cdots & 0 \\ 0 & R_2(t) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & R_n(t) \end{pmatrix},$$

for the diagonal matrix with the $R_i(t)$ along the diagonal. We claim that the processes R, $X=RX_{\mathrm{a}}$ and $Y=R\cdot Y_{\mathrm{a}}$ are all fully observed. Clearly, R is observed; RX_{a} (the ordinary matrix product of these processes) is observed since, whenever X_{a} is unobserved, $R=0$. Recall that $R\cdot Y_{\mathrm{a}}$ is the transform of Y_{a} by R, and is defined by

$$(R\cdot Y_{\mathrm{a}})(t)=\sum_{s=1}^{t} R(t)\,\Delta Y_{\mathrm{a}}(t).$$

So $R\cdot Y_{\mathrm{a}}$ is the process Y_{a} whose individual elements are *stopped*, i.e. held constant, after the time T_i of their last observations. Hence this process, also, is observable. We denote the history of the observed data X, Y and R as

$$\mathcal{F}_t=\{X(s),Y(s),R(s): s=1,\ldots,t\}$$

and define $\mathcal{F}_{t-}=\mathcal{F}_{t-1}\cup\{X(t),R(t)\}$. The following model is induced for the observed longitudinal responses Y:

$$\mathbf{E}\{\Delta Y(t)|\mathcal{F}_{t-}\}=X(t)\,\beta(t)-\mathbf{E}\{\varepsilon(t-1)|\mathcal{F}_{t-}\} \tag{11}$$

where $\varepsilon=R\cdot\varepsilon_{\mathrm{a}}$. This equality may be derived directly from the linear model for the local characteristics of Y_{a}, the fact that R is predictable and the independent censoring assumption. The

key point is that the same parameters β appear in the local characteristics of both Y and Y_a, and hence are estimable from observed data. These parameters represent the effects of covariates on the expected change in hypothetical longitudinal response at a given time and so will often have scientific relevance. In Section 4.2 we demonstrate how to estimate these parameters.

4.2. Model fitting

4.2.1. Estimation

To estimate $\beta = (\beta_1, \ldots, \beta_p)'$ we seek a matrix-valued process X^- having the property that $X^- X = I$. However, owing to drop-out such a process does not always exist. Let $\mathcal{T} = \{t : \det\{X'(t)\ X(t)\} \neq 0\}$, the set of times t at which the matrix $X'(t)\ X(t)$ is invertible. This $\mathcal{T}$ is a random set over which estimation may be reasonably undertaken, often an interval whose upper end point is reached only when very few subjects remain under observation. On $\mathcal{T}$ the matrix $\{X'(t)\ X(t)\}^{-1} X'(t)$ exists, making the process X^- given by

$$X^-(t) = \begin{cases} \{X'(t)\ X(t)\}^{-1} X'(t) & t \in \mathcal{T}, \\ 0 & t \notin \mathcal{T} \end{cases}$$

well defined. So on $\mathcal{T}$ our estimate

$$\hat{\beta}(t) = X^-(t)\{Y(t) - Y(t-1)\}$$

of $\beta(t)$ is just the ordinary least squares (OLS) estimate of this parameter, based on all available increments. Outside $\mathcal{T}$ we simply have $\hat{\beta}(t) = 0$. This leads to the estimator $\hat{B}$ of B that is given by

$$\begin{aligned} \hat{B}(t) &= \sum_{s=1}^{t} \hat{\beta}(s) \\ &= \sum_{s=1}^{t} X^-(s)\{Y(s) - Y(s-1)\} = (X^- \cdot Y)(t). \end{aligned} \tag{12}$$

Thus we set $\hat{B} = X^- \cdot Y$, the transform of Y by X^-. So defined, $\hat{B}$ is an estimator of B on $\mathcal{T}$; specifically, it estimates $B^{\mathcal{T}} = \mathbf{1}_{\mathcal{T}} \cdot B$, and there may be some small bias in estimating B. Estimation of $B^{\mathcal{T}}$ is reasonable in the present context of varying sample sizes and covariates, and is, in fact, all that can be expected of a non-parametric technique. Without parametric interpolation, there may be time points about which the data can say nothing.

This estimator is again due to Aalen (1989) in the setting of event history analysis, and to Fosen *et al.* (2006b) for more general continuous time processes. It is straightforward to show that $\hat{\beta}(t)$ is unbiased for $\mathbf{1}_{\mathcal{T}}(t)\ \beta(t)$:

$$\begin{aligned} \mathbf{E}\{\hat{\beta}(t) - \mathbf{1}_{\mathcal{T}}(t)\ \beta(t)\} &= \mathbf{E}\{X^-(t)\ \Delta Y(t) - \mathbf{1}_{\mathcal{T}}(t)\ \beta(t)\} \\ &= \mathbf{E}[X^-(t)\ \mathbf{E}\{\Delta Y(t) | \mathcal{F}_{t-}\} - \mathbf{1}_{\mathcal{T}}(t)\ \beta(t)] \\ &= \mathbf{E}\{X^-(t)[X(t)\ \beta(t) - \mathbf{E}\{\varepsilon(t-1) | \mathcal{F}_{t-}\}] - \mathbf{1}_{\mathcal{T}}(t)\ \beta(t)\} \\ &= \mathbf{E}\{\mathbf{1}_{\mathcal{T}}(t)\ \beta(t)\} - \mathbf{E}\{\varepsilon(t-1)\} - \mathbf{E}\{\mathbf{1}_{\mathcal{T}}(t)\ \beta(t)\} = 0. \end{aligned}$$

Therefore, $\hat{B}$ is unbiased for $B^{\mathcal{T}}$. What we have done is to mimic Aalen's unbiased estimator, and to show that measurement error does not affect this unbiasedness.

The estimator $\hat{B}$ is essentially a moment-based estimator of B. It sums the least squares estimates of β based on the observed increments. Crucially, nowhere do we require Y and R to be independent. We rely on an assumption that hypothetical random effects are martingales, and if this assumption breaks down then so does unbiasedness. Each surviving subject is thought to

have a mean 0 step in their random effects; non-zero expected increments in the random effects cannot be distinguished from a change in population mean.

4.2.2. Inference

Inference is discussed in Farewell (2006). Estimators of the finite sample and asymptotic variances of $\hat{B}$ are not so readily derived as in the corresponding theory of event history analysis. Counting processes behave locally like Poisson processes (Andersen *et al.*, 1992), having equal mean and variance, but this result does not hold in generality. Moreover, error ε_{a} in the measurement of the hypothetical variable leads to negatively correlated increments in $\hat{B}$ and results in a complex pattern of variability. However, computing time occupied by parameter estimation is negligible, so we recommend the use of the bootstrap for inference about B. Farewell (2006) provides a result that $\hat{B}$ is $\sqrt{n}$ consistent for B with a Gaussian limiting distribution. He also gives an approximation that, in the absence of measurement error, justifies a simple calculation using OLS regression, as outlined in Appendix A. In the application to follow, we use the bootstrap distribution for $\hat{B}$.

4.3. Diagnostics

Most diagnostic tools are based in some way on the estimated residuals from a fitted model. In the current setting the residuals are $Z = M + \varepsilon$ and may be estimated by

$$\hat{Z} = (I - H) \cdot Y^{\mathcal{T}},$$

where $H = XX^{-}$ is the hat matrix of OLS and $Y^{\mathcal{T}} = \mathbf{1}_{\mathcal{T}} \cdot Y$. Standard residual plots, e.g. of $\hat{Z}$ against fitted values or covariates, should reveal systematic misspecifications of the model for the mean response but need not show the usual random scatter since we do not assume homogeneity of variances, either between or within subjects.

One simple diagnostic that is tailored to the martingale assumption is a scatterplot of increments in the residuals, $\hat{Z}(t) - \hat{Z}(t-1)$, against $\hat{Z}(t-1)$. In the absence of measurement error, a plot of this kind should show no relationship. Substantial measurement error would induce a negative association, in which case the fit would be improved by including $\hat{Z}(t-1)$ as a covariate at time t.

We also propose two new diagnostic tools, as follows. The first is a graphical check of the martingale structure of the random effects and exploits the fact that, for $t > 1$,

$$\mathrm{cov}\{M_{\mathrm{a}}(1) + \varepsilon_{\mathrm{a}}(1), M_{\mathrm{a}}(t) + \varepsilon_{\mathrm{a}}(t)\} = \mathbf{V}\{M_{\mathrm{a}}(1)\}. \tag{13}$$

This result is easily proved, since martingales have uncorrelated increments and the errors ε are mutually independent. The point about equation (13) is that the empirical version of the left-hand side can be evaluated at each measurement time, whereas the expression on the right-hand side shows that the corresponding theoretical quantity is constant over time. Hence, a plot of $\mathrm{cov}\{\hat{Z}(1), \hat{Z}(t)\}$ against t has diagnostic value, with departures from a straight line with zero slope indicating unsuitability of model (11).

Clearly, similar plots can be derived based on the observation that

$$\mathrm{cov}\{M_{\mathrm{a}}(s) + \varepsilon_{\mathrm{a}}(s), M_{\mathrm{a}}(t) + \varepsilon_{\mathrm{a}}(t)\} = \mathbf{V}\{M_{\mathrm{a}}(s)\}$$

for all $1 \leqslant s < t$, where the above diagnostic corresponds to choosing $s = 1$. What is less clear is how much additional information is provided by such plots, since the plots are closely related.

We supplement this covariance diagnostic plot with an informal test statistic. Writing $\hat{Z}(\tau)$ for the final value that is assumed by the process $\hat{Z}$, we have in particular that

$$\mathbf{E}\{\hat{Z}'(1)\,\hat{Z}(2)\} = \mathbf{E}\{\hat{Z}'(1)\,\hat{Z}(\tau)\}.$$

Therefore $\mathbf{E}[\hat{Z}'(1)\{\hat{Z}(\tau) - \hat{Z}(2)\}] = 0$, and for large n the approximation

$$\frac{\hat{Z}'(1)\{\hat{Z}(\tau) - \hat{Z}(2)\}}{\sqrt{\mathbf{V}[\hat{Z}'(1)\{\hat{Z}(\tau) - \hat{Z}(2)\}]}} \sim N(0,1) \tag{14}$$

holds. Large absolute values of this statistic constitute evidence against the martingale hypothesis. In practice, we use the bootstrap variance in place of its theoretical equivalent in the denominator.

4.4. Summarizing remarks

In summary, our model is

$$Y_{\mathrm{a}}(t) = (X_{\mathrm{a}} \cdot B)(t) + M_{\mathrm{a}}(t) + \varepsilon_{\mathrm{a}}(t)$$

for $t = 1, \ldots, \tau$. The observed data are R, $X = RX_{\mathrm{a}}$ and $Y = R \cdot Y_{\mathrm{a}}$. We assume that

$$\mathbf{E}\{M_{\mathrm{a}}(t) | X_{\mathrm{a}}(t), R(t); X_{\mathrm{a}}(s), R(s), Y_{\mathrm{a}}(s), \varepsilon_{\mathrm{a}}(s) : s = 1, \ldots, t-1\} = M_{\mathrm{a}}(t-1)$$

and our estimator for B is

$$\hat{B}(t) = \sum_{s=1}^{t} X^{-}(s)\{Y(s) - Y(s-1)\}.$$

Appendix A illustrates how this can be implemented by using standard statistical software.

5. Simple example revisited

For further discussion we return to the simple two-time-point example that was used in Sections 2 and 3. Mixing the notation of the previous sections, our hypothetical longitudinal model can formally be expressed as

$$\mathbf{E}(Y_1) = \mu_1,$$
$$\mathbf{E}(Y_{2\mathrm{a}} - Y_1 | Y_1, \varepsilon_1) = \mu_{2\mathrm{a}} - \mu_1 - \varepsilon_1,$$

and the independent censoring assumption asserts that

$$\mathbf{E}(Y_{2\mathrm{a}} - Y_1 | Y_1, \varepsilon_1, R) = \mathbf{E}(Y_{2\mathrm{a}} - Y_1 | Y_1, \varepsilon_1).$$

Written using more traditional modelling notation, these assumptions are satisfied if

$$Y_1 = \mu_1 + M_1 + \varepsilon_1, \tag{15}$$

$$Y_{2\mathrm{a}} = \mu_{2\mathrm{a}} + M_{2\mathrm{a}} + \varepsilon_{2\mathrm{a}}, \tag{16}$$

$$\{(M_1, M_{2\mathrm{a}}), \varepsilon_1, \varepsilon_{2\mathrm{a}}\} \text{ mutually independent with zero means} \tag{17}$$

and

$$\mathbf{E}(M_{2\mathrm{a}} - M_1 | M_1, R = 1) = 0. \tag{18}$$

Under assumptions (15)–(18), our least squares estimator (12) is given by

$$\begin{aligned}\hat{\mu}_{2a} &= \bar{Y}_1 + \bar{Y}_{2a}^{\mathcal{C}} - \bar{Y}_1^{\mathcal{C}} \\ &= \bar{Y}_{2a}^{\mathcal{C}} - (\bar{Y}_1^{\mathcal{C}} - \bar{Y}_1)\end{aligned} \tag{19}$$

and is unbiased for μ_{2a}.

Consider now the assumptions that lead to the unbiasedness of $\hat{\mu}_{2a}$. Equation (15) is unremarkable; equation (16) is for the possibly counterfactual drop-out-free response Y_{2a}, as we have argued for objective 3. The zero-mean assumptions in condition (17) are needed to give μ_1 and μ_{2a} interpretations as drop-out-free population means, which are the parameters of interest. Note, though, that we do not require M_1 and M_{2a} to be independent. Equation (18) provides our key assumption, that the subject-specific random effects have mean 0 increments, conditional on that subject's observed history. It is this assumption that we test with our diagnostic in Section 4.3. An untestable consequence of equation (18), taken together with condition (17), is that the subject-specific random effects also have mean 0 increments conditional on dropping out.

Equations (15)–(18) completely specify the model and it is perhaps worth restating what has *not* been assumed. There are no distributional statements about either the random effects or the measurement errors, and there is no assumption of identical distributions across subjects. There are no statements whatsoever about Y_{2b}, what happens after drop-out. Importantly, we have not made any further assumptions on the drop-out probability $\pi(\cdot)$. This does not mean that $\pi(\cdot)$ is entirely unrestricted: condition (18) holds if, and only if,

$$\mathbf{E}[\Delta\{1 - \pi(M_1, \Delta)\}|M_1] = 0, \tag{20}$$

where $\Delta = M_{2a} - M_1$. Examples that satisfy the above condition include a random-intercept model in which $\Delta = 0$, with any $\pi(\cdot)$, an independent censoring drop-out model in which $\pi(M_1, \Delta) = \pi(M_1)$, with any Δ for which $\mathbf{E}(\Delta|M_1) = 0$, and any $\pi(M_1, \Delta)$ that is an even function of Δ, taken together with any zero-mean, symmetric distribution $[\Delta|M_1]$.

None of these examples are drop-out MAR models, since in every case $\pi(Y_1, Y_{2a}) \neq \pi(Y_1)$. Notwithstanding this comment, in the first two examples we have drop-out probability depending only on the most recent random effect M_1. In this sense our assumptions are similar to sequential drop-out MAR (Hogan *et al.*, 2004), with the additional assumption of martingale random effects. Nevertheless, and as the third example illustrates, it is possible to construct a variety of models for which $\pi(M_1, \Delta) \neq \pi(M_1)$ yet condition (20) remains true.

6. Simulations

We demonstrate the use of the covariance diagnostics in two simulation studies. Pitting a martingale random-effects process against a popular non-martingale alternative, we report the estimated power and type I error rates of the informal test (14) and illustrate the suggested covariance plots.

6.1. Scenario 1

The first simulation scenario mimics the schizophrenia example that is to be considered in Section 7, though with just one treatment group and so no covariates. Measurements are scheduled at weeks $(w_1, \ldots, w_6) = (0, 1, 2, 4, 6, 8)$.

Let $U_0, U_1, U_2, \ldots$ be independent mean 0 Gaussian n-vectors, which we use to construct two random-effects processes. Put $S_a(0) = M_a(0) = 0$, and for non-negative t define

$$\begin{aligned}S_a(t) &= U_0 + U_1 w_t, \\ M_a(t) &= U_0 + U_1 + U_2 + \ldots + U_{t-1}.\end{aligned}$$

Then S_a is a random-intercept and slope process, of the kind that was described by Laird and Ware (1982), whereas M_a is a martingale. We take $\mathbf{V}(U_0) = \sigma_0^2 I$ and $\mathbf{V}(U_1) = \sigma_1^2 I$ and choose the variances of the further values to ensure that $\mathbf{V}\{S_a(t)\} = \mathbf{V}\{M_a(t)\}$. This set-up allows us to compare these two types of random-effects process with, as far as is possible, all else being equal.

The responses are now defined as

$$Y_a^S(t) = \mu_t + S_a(t) + \varepsilon_a(t),$$
$$Y_a^M(t) = \mu_t + M_a(t) + \varepsilon_a(t),$$

with $\varepsilon_a(t) \sim N(0, \sigma_\varepsilon^2 I)$, and independence between time points. The probabilities of drop-out between times t and $t+1$ are logistic with exponents $\alpha_t + \gamma_t S_a(t)$ and $\alpha_t + \gamma_t M_a(t)$ for Y^S and Y^M respectively.

For each of n = 125, 250, 500, 1000 we took 1000 simulations from this model. We used $\mu_1 = \ldots = \mu_6 = 0$ and chose the other parameter values to correspond roughly to the schizophrenia data: $\sigma_0^2 = 200$, $\sigma_1^2 = 15$, $\sigma_\varepsilon^2 = 100$ and

$$(\alpha_1, \ldots, \alpha_5) = (-8, -6, -6, -6, -4),$$
$$(\gamma_1, \ldots, \gamma_5) = (0.2, 0.3, 0.3, 0.5, 0.6).$$

This led to about 50% drop-out in each model, spread over time points 2–5, with only about 1% of subjects dropping out after just one observation. Each data set was analysed by using our linear increments (LI) approach, an IPW estimating equation approach and by fitting a multivariate normal distribution with unstructured within-subject covariance matrix (method UMN). Under both the IPW and UMN methods we made a misspecified drop-out MAR assumption. For IPW we used response at time $t-1$ as covariate in a logistic model for drop-out at time t. No drop-out model is needed for UMN under drop-out MAR.

Table 1 summarizes results at $n = 500$. There was severe downward bias in the observed mean values (OLS) for each of Y_a^M and Y_a^S and this is only partly corrected by the misspecified IPW or UMN methods. The LI fit to Y^M shows no bias, as expected, and confidence interval coverage is good. The observed mean bias was improved but not removed when our method is used on Y^S, unsurprisingly given that the model is then also misspecified. Usually such misspecification would be detected by the diagnostics. For example, box plots of the residual covariances (Fig. 1) suggest good diagnostic power for distinguishing the models and this is confirmed by the performance of the test statistic (14), for the variance of which we used 100 bootstrap samples for each data set (Table 2).

6.2. Scenario 2

For the next simulation we introduce covariates and change the drop-out model. As well as an intercept term we include a time constant Bernoulli(0.5) covariate and also a time-varying covariate, independently distributed as $N(0, \sigma_W^2)$ at each time point. In the notation of Section 4, the corresponding cumulative regression functions are taken to be

$$B(t) = (0, \mathbf{1}(t > 0) \exp\{-(t-1)\}, t)'.$$

We add to the mix some error in measurement ε, arising according to a t-distribution on ν degrees of freedom and scaled by a factor σ_ε, i.e. $\sigma_\varepsilon^{-1} \varepsilon_i(t) \sim t(\nu)$. The final measurement times

Table 1. Estimated mean responses and standard errors SE for scenario 1 using observed data without correction for drop-out (OLS), with IPW or a multivariate normal model with unstructured covariance matrix (UMN), both of which falsely assume that drop-out is MAR, and under the LI method†

Method			*Results for the following value of w:*					
			0	*1*	*2*	*4*	*6*	*8*
Y^M	OLS	Mean	0.00	−0.30	−2.75	−4.34	−10.61	−19.41
		SE	0.77	0.78	0.77	0.91	1.32	1.89
	IPW	Mean	−0.03	−0.03	−1.12	−2.25	−6.10	−13.17
		SE	0.78	0.81	0.84	1.12	2.04	2.83
	UMN	Mean	−0.02	−0.02	−0.53	−1.80	−6.00	−12.90
		SE	0.77	0.77	0.85	0.91	1.44	1.83
	LI	Mean	0.00	−0.02	−0.02	0.01	0.05	0.02
		SE	0.77	0.78	0.89	0.97	1.55	2.05
		Cov (%)	96.4	94.1	95.2	94.3	94.8	94.6
Y^S	OLS	Mean	−0.01	0.26	−2.90	−5.08	−12.95	−22.38
		SE	0.79	0.82	0.83	1.06	1.11	1.34
	IPW	Mean	0.01	−0.17	−1.25	−2.84	−8.06	−15.67
		SE	0.79	0.82	0.97	1.16	1.68	1.83
	UMN	Mean	0.01	−0.15	−0.75	−2.38	−7.12	−13.45
		SE	0.79	0.82	0.89	1.12	1.16	1.39
	LI	Mean	−0.01	0.02	−0.16	−0.98	−3.61	−7.81
		SE	0.79	0.82	0.93	1.20	1.18	1.44
		Cov (%)	94.8	95.7	94.1	85.9	19.8	0.1

†The coverage Cov of nominal 95% confidence intervals under LI is also included. The sample size was $n = 500$, and results were averaged over 1000 simulations.

$T_1, \ldots, T_n$ are determined by the relationship

$$\text{logit}\{\mathbf{P}(T = t | T \geqslant t, U_0, \ldots, U_{t-1})\} = \begin{cases} -\infty & t = 0, \\ \alpha + S_a(t) + M_a(t) & t = 1, \ldots, 6, \\ \infty & t = 7, \end{cases}$$

so that $1 \leqslant T_i \leqslant 7$ for each i.

We defined

$$Y_a^S = X_a \cdot B + S_a + \varepsilon_a,$$
$$Y_a^M = X_a \cdot B + M_a + \varepsilon_a.$$

The parameters were taken to be $\sigma_0 = \sigma_1 = \sigma_W = 1, \pi = \frac{1}{2}, \sigma_\varepsilon = \frac{1}{3}, \nu = 3$ and $\alpha = -7$. This gave approximately 25% drop-out, roughly evenly spread over times 2–6. Again 100 bootstrap samples were drawn to compute variances for the test statistic (14).

Mean estimates of B for sample size $n = 500$ using both Y^M and Y^S are shown in Fig. 2, together with the true values and ± 2 empirical standard errors around the Y^M-estimates. Bootstrap standard errors matched the empirical values closely. Standard errors derived from asymptotic results, which avoid the need to bootstrap but at the expense of assuming negligible measurement error, were slightly conservative, overestimating typically by about 5%. As expected there was no evidence of bias for our increment-based estimates of B based on Y^M. Estimates from

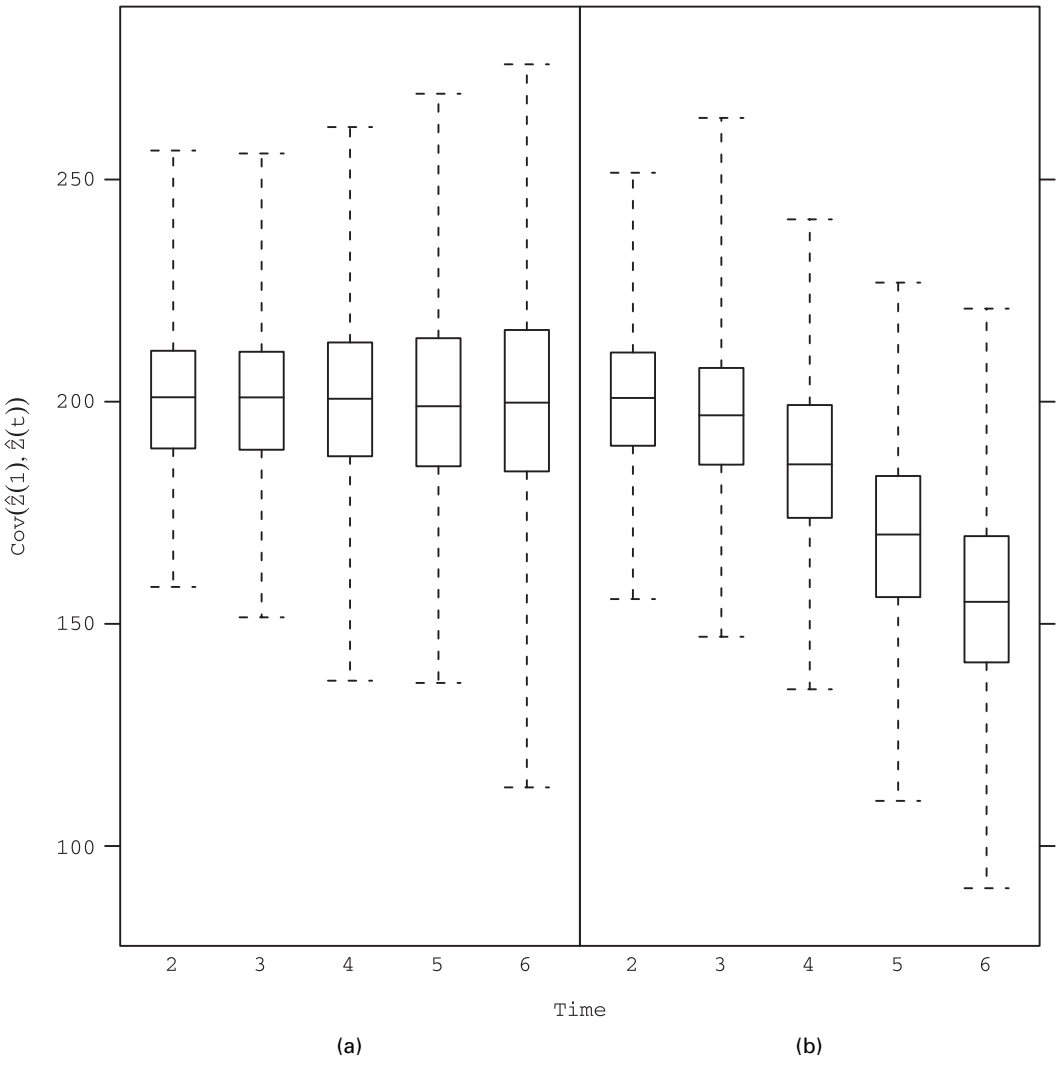

Fig. 1. Box plots of cov$\{\hat{Z}(1), \hat{Z}(t)\}$ based on 1000 simulations under scenario 1 at sample size $n = 500$: (a) true martingale structure Y^M; (b) Laird–Ware random-intercept and slope structure Y^S

Table 2. Estimated size and power of the diagnostic test, based on simulation results

Scenario		*Results for the following values of n:*			
		125	*250*	*500*	*1000*
1	Power	0.307	0.530	0.766	0.980
	Type I error	0.056	0.056	0.053	0.059
2	Power	0.147	0.241	0.390	0.686
	Type I error	0.056	0.059	0.045	0.052

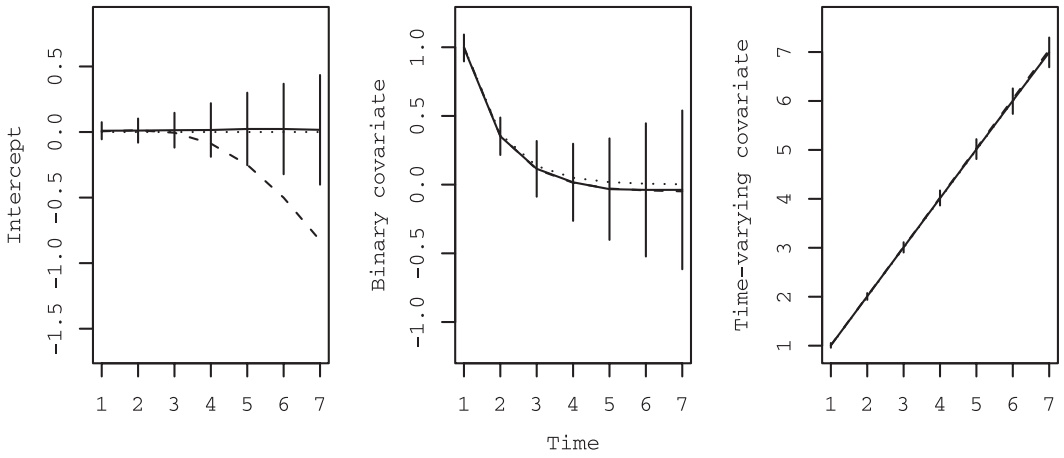

Fig. 2. Summary of estimates $\hat{B}$ for scenario 2, at sample size $n = 500$: mean dynamic estimates from Y^M (———) and Y^S together with true values (·····)

the misspecified model for Y^S were also good for B_2 and B_3; in fact so close that the lines in the plots are hardly distinguishable. There was, however, bias for the intercept B_1. Identification of the random-effect structure through residual covariances was more difficult than for scenario 1, causing some loss of power for the test statistic (Table 2).

7. Analysing data from a longitudinal trial

We now describe an application of the methods of Section 4 to data from the schizophrenia clinical trial that was introduced earlier. The trial compared three treatments: a placebo, a standard therapy and an experimental therapy. The response of interest, PANSS, is an integer ranging from 30 to 210, where high values indicate more severe symptoms. A patient with schizophrenia entering a clinical trial may typically expect to score around 90.

Of the 518 participants, 249 did not complete the trial, among whom 66 dropped out for reasons that were unrelated to their underlying condition. The remaining 183 represent potentially informative drop-out, though we emphasize that our new approach does not need to distinguish these from the non-informative drop-outs. We mention them only because we shall refer to other procedures that draw such a distinction.

The goal of the study was to compare the three treatments with respect to their ability to improve (reduce) the mean PANSS-score. The patients were observed at base-line ($t = 1$) and thereafter at weeks 1, 2, 4, 6 and 8 ($t = 2, 3, 4, 5, 6$) of the study. The only covariates used here are treatment groups. The dotted curves in Fig. 3 show for reference the observed mean response at each time in each treatment group, calculated in each case from subjects who have not yet dropped out. Hence, the plotted means estimate conditional expectations of the PANSS-score (objective 2), which are not necessarily the appropriate targets for inference.

Fig. 3 displays the pronounced differences between the OLS estimates and their dynamic linear counterparts. The OLS estimates invite the counter-intuitive conclusion that, irrespective of treatment type, patients' PANSS-scores decrease (improve) over time. By contrast, our increment-based estimator suggests that this is a feature of informative drop-out, and that patients on the placebo do not improve over time; in fact, there is even a suggestion that their PANSS-scores increase slightly. The levelling out of treatment effects over time that is seen under our new approach is also unsurprising.

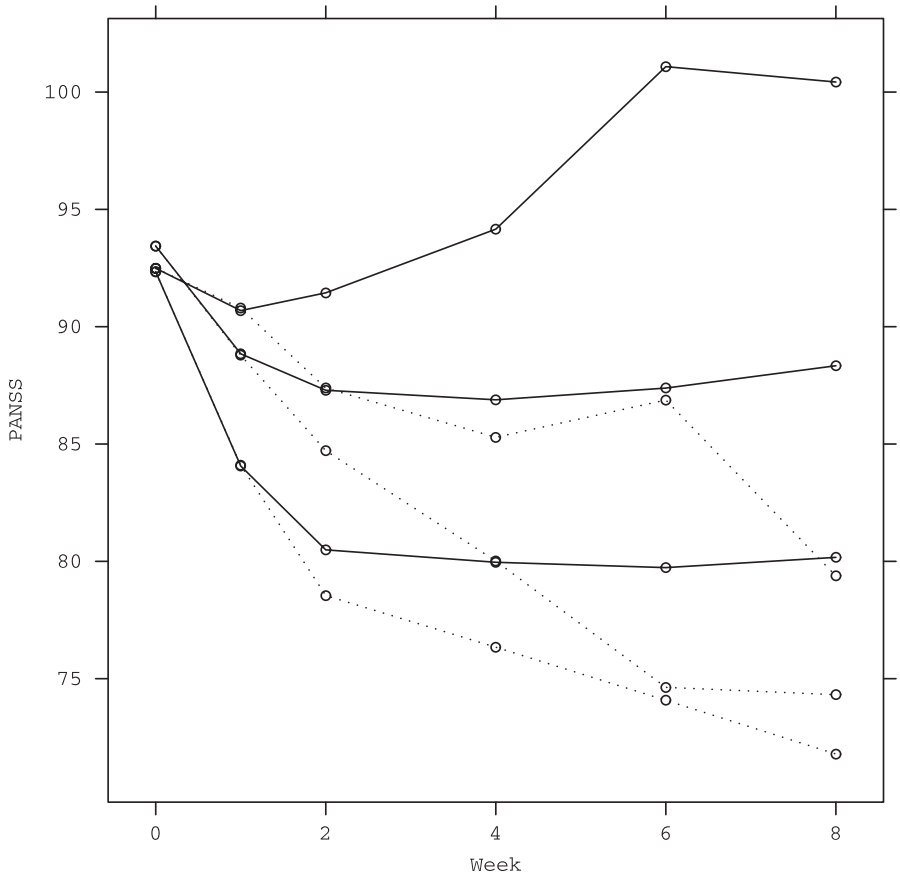

Fig. 3. Estimated PANSS mean values under OLS (·····) and our dynamic linear approach (———): the topmost curves correspond to the placebo group, the middle curves to the standard treatment group and the lowest curves to the experimental treatment group

In Fig. 4 and Table 3 we compare the dynamic linear fits with those which were obtained under four other approaches. Fig. 4 shows the estimated means for each treatment group whereas Table 3 gives for standard treatment the estimated mean change in response between the beginning and end of the study, together with the effect of placebo or experimental treatment on this quantity. The other approaches are as follows:

(a) maximum likelihood estimation under a multivariate normal model with unstructured covariance matrix (method UMN) (this approach assumes that drop-out is MAR);
(b) a quadratic random-effects joint longitudinal and event time informative drop-out model that was fitted by Dobson and Henderson (2003) using EM estimation, as suggested by Wulfsohn and Tsiatis (1997) (Dobson and Henderson compared four random-effects structures and concluded that, between these, the model that is used here with random-intercept, slope and quadratic terms 'is strongly preferred by likelihood criteria, even after penalizing for complexity';
(c) an IPW estimating approach as described by Robins *et al.* (1995), with a logistic drop-out MAR model.
(d) a second martingale fit (DYN) in which residuals at time t are included as covariates for the increments between t and $t+1$, along the lines of the dynamic covariate approaches

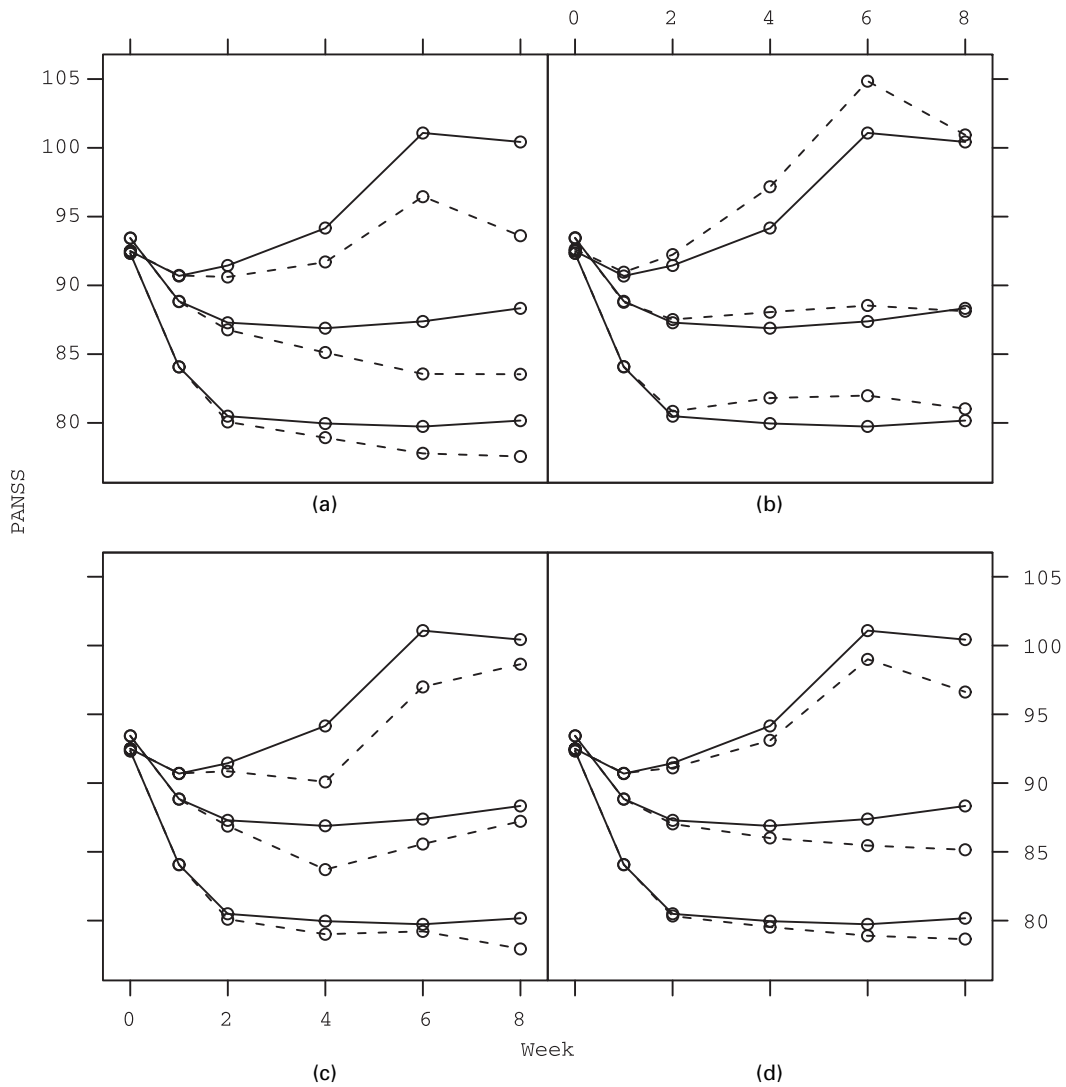

Fig. 4. Estimated PANSS mean values for (from top to bottom pairs of curves, in every case) the placebo, standard and experimental groups (- - - - -, estimates generated under methods (a)–(d) in the text; ———, estimates under the dynamic linear approach): (a) method UMN; (b) Dobson and Henderson's (2003) method; (c) IPW method; (d) method DYN

for event history analyses that were described by Aalen *et al.* (2004) and Fosen *et al.* (2006a).

There are broad similarities between our increment-based estimates and any of approaches (a)–(d) but some differences are worth noting. Method (a) gives a smaller adjustment to the observed means than the others, whereas method (c) adjusts almost as much as our linear increment fits. Both of these are drop-out MAR models. Method (b) assumes a Gaussian response but method (c) has no modelling assumptions for the responses, a gain that is obtained at the expense of an increase in standard errors. Method (d) leads to estimates that are comparable with the fit that is obtained by using only exogenous covariates, albeit slightly closer to the

Table 3. Effect of treatment on change in mean response (week 8 minus week 0) under the LI approach (12), OLS with an independence assumption and methods (a)–(d) described in the text†

Treatment	*Results for the following methods:*					
	LI	*OLS*	*(a) (UMN)*	*(b) (Dobson and Henderson, 2003)*	*(c) (IPW)*	*(d) (DYN)*
S	−5.10	−19.12	−9.90	−5.34	−6.22	−8.29
	(3.49)	(3.43)	(3.06)	(2.94)	(7.72)	(3.21)
P – S	13.04	6.01	11.01	13.66	12.37	12.42
	(5.32)	(5.01)	(4.49)	(5.29)	(8.82)	(4.82)
E – S	−7.07	−1.43	−4.89	−5.97	−8.18	−5.40
	(3.80)	(3.86)	(3.38)	(3.37)	(7.83)	(3.73)

†'S' represents the standard treatment, 'P' placebo and 'E' the experimental treatment. Standard errors are in parentheses.

observed means. Method (b), the quadratic random-effects model, gives estimates that are close to those obtained by using our new approach. Method (b) took several days of computing time to fit, whereas estimates for other models can be obtained quickly, our linear increment models in particular. The availability of a closed form estimator (12) meant that the 1000 bootstrap simulations that were needed to compute the standard errors were completed in under 10 on an unremarkable laptop computer. In Appendix A, we demonstrate briefly one way in which our dynamic linear models may be implemented by using standard software.

It is interesting to recall that, in approach (b), Dobson and Henderson (2003) modelled the drop-out process explicitly and distinguished censoring due to inadequate response from other censoring events; neither is necessary under our proposed approach. Given the similarities between our dynamic linear results and those of method (b), the Dobson and Henderson assumption that these other events are uninformative about PANSS seems to be justified.

The diagnostics proposed may be illustrated by using these data. Having computed $\hat{B}$, it is straightforward to extract $\hat{Z}$. Fig. 5 shows $\hat{Z}(t) - \hat{Z}(t-1)$ against $\hat{Z}(t-1)$ at each time point and provides some evidence that our original model is misspecified. Fig. 5(a) for week 1 clearly indicates a weak negative association, which is consistent with measurement error in the response. The effect is less marked in later weeks. As discussed in Section 4.3, this suggests considering inclusion of $\hat{Z}(t-1)$ as an additional covariate in the model for increments at time t, which is approach (d) above. Fig. 4 (d) shows that the fitted mean response profiles are not materially affected by the misspecification that is indicated by Fig. 5.

Box plots illustrating the bootstrap distribution of the diagnostic $n^{-1}\hat{Z}'(1)\hat{Z}(t)$ are shown in Fig. 6. The plot includes results for $t = 1$ to exhibit the magnitude of the independent noise terms. Since the covariance is expected to be constant only for $t > 1$, for diagnostic purposes the first box plot may be safely ignored. On the basis of remaining box plots, derived from 1000 bootstrap samples, there is evidence of a downward trend in the diagnostic. However, this is mild, and the informal test statistic (again based on 1000 bootstrap samples) is -1.61, corresponding to a p-value of about 0.1. Together, the diagnostics suggest that departures from the model are sufficiently small to be of little concern.

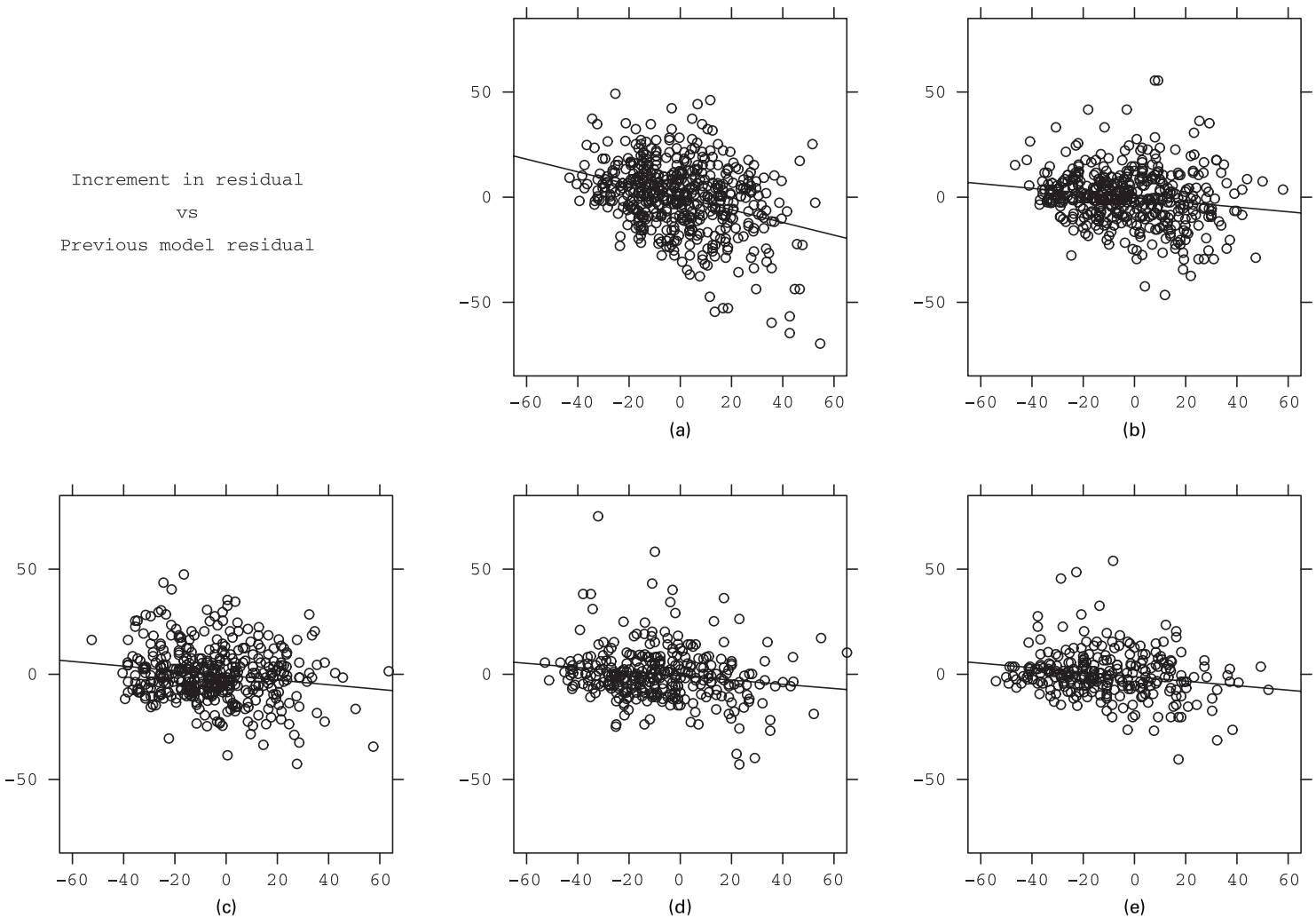

Fig. 5. PANSS data: residual increments $\hat{Z}(t) - \hat{Z}(t-1)$ plotted against $\hat{B}(t-1)$: (a) week 1; (b) week 2; (c) week 4; (d) week 6; (e) week 8

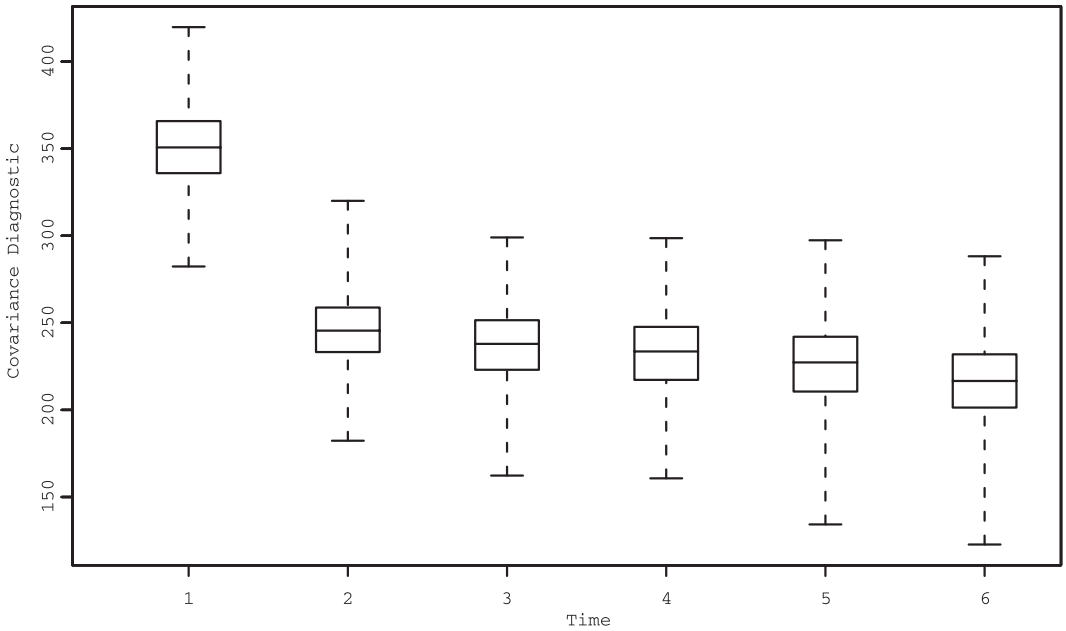

Fig. 6. PANSS data: box plots of $\text{cov}\{\hat{Z}(1), \hat{Z}(t)\}$ from 1000 bootstrap samples: for a correctly specified model the mean values for $t > 1$ should be equal

8. Discussion

Many approaches to the analysis of longitudinal data with drop-out begin with the idea of vectors of complete data Y, observed data Y_{obs} and missingness indicators R. We have argued that this set-up can be too simple, as it does not recognize that drop-out can be an event that occurs in the lives of the subjects under study and that can affect future responses. Distributions after drop-out may be different from those that would have occurred in the absence of that event, an extreme example being when drop-out is due to death. Another might be when drop-out is equivalent to discontinuing a treatment. Thus there is no well-defined complete-data vector Y and we are led into the world of counterfactuals, as described for the two-time-point example of Section 2, and the need for careful thought about objectives and targets for inference. An exception is when inference is conditional on drop-out time (objective 2) and hence based only on observed data. Otherwise, untestable assumptions of one form or another are required for inference. In this paper we consider interest to lie in the drop-out-free response Y_{a} and make the two key assumptions of independent censoring and martingale random effects.

In our view, the analysis of longitudinal data, particularly when subject to missingness, should always take into account the time ordering of the underlying longitudinal processes. Often, the drop-out decision is made between measurement times, and we acknowledge this by insisting that the drop-out process be predictable, while allowing it to depend arbitrarily on the past. Subsequent events could be affected by the drop-out decision, and in this sense drop-out could be informative about future longitudinal responses. We reiterate that we do not require *all* future values to be independent of the drop-out decision: the realized response is free to depend on this decision. Nor is the required independence unconditional: our assumption is that, given everything that has been observed, drop-out status gives no new information about the mean of the next *hypothetical* response. This is a weaker and, to us, more logical assumption than the standard drop-out MAR form. Ultimately, however, both the drop-out MAR and the independent

censoring assumption share the same purpose: to enable inference by making assumptions about the drop-out process. Drop-out MAR enables inference using the observed data likelihood, whereas independent censoring enables inference using the observed local characteristics.

What *is* therefore important is that all relevant information in $\mathcal{F}_t$ should be included in the model for the next expected increment. For example, Fig. 5 suggested inclusion of the previously observed residual as a covariate for current increments. A similar approach might be used to simplify variance estimation, or if there are subject-specific trends, as in a random-slope model. Aalen *et al.* (2004) advocated an equivalent approach in dynamic linear modelling of recurrent event data. We note also the argument in Fosen *et al.* (2006a) that use of residuals $\hat{Z}$ rather than Y helps to preserve the interpretation of exogenous covariate effects.

Modelling the local characteristics acknowledges the time ordering in longitudinal data analysis, naturally accounting for within-subject correlation and possibly history-dependent drop-out. These features can all be accommodated through linear models on the observed increments of the response process. At no great loss of understanding, the applied statistician could think of our procedure as 'doing least squares on the observed response increments, then accumulating', to draw inference about the longitudinal features that a population would have exhibited, assuming that no-one had dropped out.

Thus far, we have assumed a balanced study design, by which we mean a common set of intended measurement times for all subjects. A natural extension is to unbalanced study designs. It would also be of interest to consider more complicated random-effects models for the increments of a longitudinal process, potentially gaining efficiency but requiring additional parametric assumptions. We have not so far explored this option; nor the important but challenging possibility of developing sensitivity procedures for our approach.

Acknowledgements

The authors are grateful for the detailed comments and helpful advice of all referees for the paper. Peter Diggle is supported by an Engineering and Physical Sciences Research Council Senior Fellowship. Daniel Farewell's research was carried out during his Medical Research Council funded studentship at Lancaster University. Robin Henderson is grateful for valuable discussions with Ørnulf Borgan and Niels Keiding at the Centre for Advanced Study, Oslo.

Appendix A: Fitting dynamic linear models by using standard software

Least squares equations may be solved, and hence our proposed models fitted, in virtually all software for statistical computing. We note, reflecting our own computing preferences, that this is particularly straightforward by using the `lmList` command from the `nlme` package (Pinheiro and Bates, 2000) in `R` or `S-PLUS`. For example, to fit the dynamic linear models of Section 4 to the schizophrenia data, we constructed a data frame `schizophrenia`, having columns `i` (a unique identifier), `time` (running from 1 to T_i for each i), `treat` (a factor indicating the treatment regime) and `PANSS`. This last column stores the *change* in PANSS that is associated with the given subject and time point, i.e. it contains $\Delta Y_i(1), \ldots, \Delta Y_i(T_i)$ for every i. Then

```
> fit <- lmList(PANSS ~ treat|time, data = schizophrenia, pool = F)
```

returns an object containing a list of estimates $\hat{\beta}(t)$ of $\beta(t)$ for each $t \in \mathcal{T}$, which may be extracted by way of the `coef` method. The cumulative sum of these estimates

```
> apply(coef(fit), 2, cumsum)
```

yields $\hat{B}$. Additionally, estimated standard errors

```
> SEs <- summary(fit)$coef[, "Std. Error", ]
```

may be extracted from the fitted model if measurement error is thought to be negligible. These estimates (squared) may be summed

```
> apply(SEs^2, 2, cumsum)
```

to yield an estimate of $\mathbf{V}(\hat{B})$ without the need for bootstrapping.

References

Aalen, O. O. (1978) Nonparametric inference for a family of counting processes. *Ann. Statist.*, **6**, 701–726.

Aalen, O. O. (1987) Dynamic modelling and causality. *Scand. Act. J.*, no. 1987, 177–190.

Aalen, O. O. (1989) A linear regression model for the analysis of life times. *Statist. Med.*, **8**, 907–925.

Aalen, O. O., Weedon-Fekær, H., Borgan, Ø. and Husebye, E. (2004) Dynamic analysis of multivariate failure time data. *Biometrics*, **60**, 764–773.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1992) *Statistical Models based on Counting Processes*. Berlin: Springer.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Berzuini, C. and Larizza, C. (1996) A unified approach for modelling longitudinal and failure time data, with application in medical monitoring. *IEEE Trans. Pattn Anal. Mach. Intell.*, **18**, 109–123.

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. R. Statist. Soc.* A, **169**, 571–584.

Copas, J. B. and Eguchi, S. (2005) Local model uncertainty and incomplete-data bias (with discussion). *J. R. Statist. Soc.* B, **67**, 459–513.

Copas, J. B. and Li, H. G. (1997) Inference for non-random samples (with discussion). *J. R. Statist. Soc.* B, **59**, 55–95.

Davidian, M., Tsiatis, A. A. and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statist. Sci.*, **20**, 261–301.

Diggle, P. J. (1993) On informative and random drop-outs in longitudinal studies. *Biometrics*, **49**, 947–949.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–93.

Dobson, A. and Henderson, R. (2003) Diagnostics for joint longitudinal and drop-out time modelling. *Biometrics*, **59**, 741–751.

Dufoil, C., Brayne, D. and Clayton, D. (2004) Analysis of longitudinal studies with death and drop-out: a case study. *Statist. Med.*, **23**, 2215–2226.

Farewell, D. M. (2006) Linear models for censored data. *PhD Thesis*. Lancaster University, Lancaster.

Fosen, J., Borgan, Ø., Weedon-Fekjaer, H. and Aalen, O. (2006a) Dynamic analysis of recurrent event data using the additive hazard model. *Biometr. J.*, **48**, 381–398.

Fosen, J., Ferkingstad, E., Borgan, Ø. and Aalen, O. O. (2006b) Dynamic path analysis—a new approach to analyzing time-dependent covariates. *Liftime Data Anal.*, **12**, 143–167.

Frangakis, C. E. and Rubin, D. B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, **86**, 365–379.

Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models. *J. R. Statist. Soc.* B, **55**, 757–796.

Heitjan, D. F. (1994) Estimation with missing data. *Biometrics*, **50**, 580.

Henderson, R., Diggle, P. J. and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

Hogan, J. W. and Laird, N. M. (1997a) Mixture models for the joint distribution of repeated measures and event times. *Statist. Med.*, **16**, 239–257.

Hogan, J. W. and Laird, N. M. (1997b) Model-based approaches to analysing incomplete longitudinal and failure time data. *Statist. Med.*, **16**, 259–272.

Hogan, J. W., Roy, J. and Korkontzelou, C. (2004) Tutorial in biostatistics — handling drop-out in longitudinal studies. *Statist. Med.*, **23**, 1455–1497.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.

Kenward, M. G. (1998) Selection and models for repeated measures with nonrandom drop-out: an illustration of sensitivity. *Statist. Med.*, **17**, 2723–2732.

Kenward, M. G., Molenberghs, G. and Thijs, H. (2003) Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 103–126.

Kurland, B. F. and Heagerty, P. J. (2005) Directly parametrized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, **6**, 241–258.
Laird, N. M. and Ware, J. H. (1982) Random effect models for longitudinal data. *Biometrics*, **38**, 963–974.
Laird, N. M. (1988) Missing data in longitudinal studies. *Statist. Med.*, **7**, 305–315.
Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Lin, D. Y. and Ying, Z. (2001) Semiparametric and nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass.*, **96**, 103–113.
Lin, D. Y. and Ying, Z. (2003) Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics*, **4**, 385–398.
Little, R. J. A. (1993) Pattern mixture models for multivariate incomplete data. *J. Am. Statist. Ass.*, **88**, 125–134.
Little, R. J. A. (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.
Little, R. J. A. (1995) Modelling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Ass.*, **90**, 1112–1121.
Little, R. J. (1998) Missing data. In *Encyclopedia of Biostatistics*, 1st edn., vol. 4 (eds P. Armitage and T. Colton), pp. 2622–2636. Chichester: Wiley.
Little, R. and Yau, L. (1996) Intention-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, **52**, 1324–1333.
Ma, G., Troxel, A. B. and Heitjan, D. F. (2005) An index of local sensitivity to nonignorable drop-out in longitudinal modelling. *Statist. Med.*, **24**, 2129–2150.
Martinussen, T. and Scheike, T. (2000) A nonparametric dynamic additive regression model for longitudinal data. *Ann. Statist.*, **28**, 1000–1025.
Martinussen, T. and Scheike, T. (2006) *Dynamic Regression Models for Survival Data*. New York: Springer.
Michiels, B., Molenberghs, G. and Lipsitz, S. R. (1999) Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, **55**, 978–983.
Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, **84**, 33–44.
Peng, Y., Little, R. J. A. and Raghunathan, T. E. (2004) An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics*, **60**, 598–607.
Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS*. New York: Springer.
Ridout, M. S. (1991) Testing for random drop-outs in repeated measurement data. *Biometrics*, **47**, 1617–1621.
Robins, J. (1998) Correction for non-compliance in equivalence trials. *Statist. Med.*, **17**, 269–302.
Robins, J. and Rotnitzky, A. (2004) Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, **91**, 763–784.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*, **90**, 106–121.
Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Statist. Ass.*, **93**, 1321–1339.
Rotnitzky, A., Scharfstein, D., Su, T.-L. and Robins, J. (2001) Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, **57**, 103–113.
Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, **47**, 1213–1234.
Rubin, D. B. (2004) Direct and indirect causal effects via potential outcomes. *Scand. J. Statist.*, **31**, 161–170.
Scharfstein, D. O., Daniels, M. J. and Robins, J. (2003) Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, **4**, 495–512.
Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Statist. Ass.*, **94**, 1096–1146.
Shih, W. J. (1992) On informative and random drop-outs in longitudinal studies. *Biometrics*, **48**, 970–972.
Sommer, A. and Zeger, S. L. (1991) On estimating efficacy from clinical trials. *Statist. Med.*, **10**, 45–52.
Troxel, A. B., Ma, G. and Heitjan, D. F. (2004) An index of local sensitivity to nonignorability. *Statist. Sin.*, **14**, 1221–1238.
Tsiatis, A. A. and Davidian, M. A. (2004) Joint modelling of longitudinal and time-to-event data: an overview. *Statist. Sin.*, **14**, 809–834.
Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M. G. (2001) Sensitivity analysis for nonrandom drop-out: a local influence approach. *Biometrics*, **57**, 7–14.
Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, **44**, 175–188.
Wulfsohn, M. S. and Tsiatis, A. A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.
Xu, J. and Zeger, S. L. (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. *Appl. Statist.*, **50**, 375–387.