Analysis of Longitudinal Data with Drop-out: Objectives, Assumptions and a Proposal

Peter Diggle
(Lancaster University and Johns Hopkins University School of Public Health)

Daniel Farewell*
(Cardiff University)

and

Robin Henderson
(University of Newcastle upon Tyne)

*Corresponding author.

Department of Epidemiology, Statistics and Public Health, Centre for Health Sciences Research, Cardiff University, Neuadd Meirionnydd, Heath Park, Cardiff, CF14 4YS.
Email: farewell@cf.ac.uk

February 19, 2007

Summary: The problem of analysing longitudinal data complicated by possibly informative drop-out has received considerable attention in the statistical literature. Most authors 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 summarise 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 prepared 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.

Key words: 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 pre-specified 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.

A useful classification of drop-out mechanisms is the hierarchy introduced by Rubin (1976) in the wider context of missing data. Dropout 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 had they not dropped out. Dropout is missing at random (MAR) if the drop-out probability may depend on observed responses but, given the observed responses, is conditionally independent of unobserved responses. Dropout is missing not at random (MNAR) if it is not MAR. Note that we interpret 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) develops 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 MAR, whereas other methods for inference, such as the original form of generalised estimating equations (Liang and Zeger, 1986), require 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 MAR is potentially inefficient; for this reason, the combination of MAR and separate parameterisation is sometimes called ignorable, and either 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 three-fold. 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, there are many possible expectations 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 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 using the two time-point scenario for clarity. Our third purpose in this article is to offer in Section 4 an approach 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. Nonetheless, all methods for dealing with missing data require, to some extent, untestable assumptions, and ours is no exception. However, we are prepared 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 prior to 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. An appendix 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 re-analyse the data from their clinical example to emphasise 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 to 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 \( \mu_t = E(Y_t) \), though similar arguments apply to the full distributions of the response variables. We emphasise 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 due 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 following model:

\[
Y_1 = \mu_1 + Z_1 \quad Y_2 = \mu_2 + Z_2 \quad E(Z_1) = E(Z_2) = 0. \tag{1}
\]

The parameter \( \mu_1 \) is the population mean at time 1. Writing down (1) invites a similar interpretation for \( \mu_2 \). In fact, the apparently straightforward adoption of (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 will 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

\[
Y_1 = \mu_1 + Z_1 \\
Y_{2a} = \mu_{2a} + Z_{2a} \\
Y_{2b} = \mu_{2b} + Z_{2b} \\
\mathbb{P}(R = 0|S) = \pi(S).
\]

In (2), \(E(Z_1) = E(Z_{2a}) = E(Z_{2b}) = 0,\) \(S\) denotes a set of conditioning variables and we allow \(\pi(\cdot)\) to depend arbitrarily on \(S.\) We make no assumption of independence between \(Z_1, Z_{2a}, Z_{2b},\) and for the unconditional case \(S = \emptyset\) we write \(\pi = \pi(\emptyset) = \mathbb{P}(R = 0).\) By construction, the parameters \(\mu_1, \mu_{2a}\) and \(\mu_{2b}\) 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, denoted group \(C.\) 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 had the subject in fact dropped out.

The drop-outs, group \(D,\) 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 had the subject continued treatment. This situation, where drop-out does not lead to cessation of measurement, is one 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 \(D.\) 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 had the subject in question 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 \(D.\) 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 needing 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 \(\mu_1 = E(Y_1)\) 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 \(E(Y_2).\) We discuss this within the specific setting of the model (2).
Objective 1: Realised second response

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

\[ Y_2 := Y_{2a}R + Y_{2b}(1 - R), \]  

which is unobserved for subjects in group \( D \). 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 zero and the definition of \( Y_2 \) above is, for practical purposes, meaningless.

In contrast, the data we analyse in Section 7 come from a longitudinal randomised 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 dropout. In this setting, \( Y_2 \) as defined at (3) is readily interpretable as the intention-to-treat response.

Objective 2: Conditional second response

A second possible target for inference is the response at time \( t = 2 \) conditional upon 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 \( C \), 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 sub-population 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 due to inadequate response to treatment.

Objective 3: Hypothetical second response

Our third potential target for inference, again unobserved for group \( D \) 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, dropout occurs if and only if \( Z_{2a} < 0 \). This might seem an extreme example, but 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 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 had they not dropped out. This is likely to be of greater value than the realised or conditional second responses, since treatment performance is of more concern than subject profiles. We emphasise, however, that this need not always be the case, and that in some circumstances a combination of objectives may be appropriate. For example, Dufouil et al. (2004) and Kurland and Heagerty (2005) separately discuss applications in which there are two causes of drop-out: death and possibly informative loss to follow-up (LTFU). In these applications the appropriate target for inference is the response distribution in the hypothetical absence of LTFU 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 (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, 1997b, 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

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

which estimates

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

3.2 Pattern mixture

A complete case analysis forms one component of a pattern mixture approach (Little, 1993), in which we formulate a separate sub-model 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 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, Molenberghs and Thijs (2003) discuss the construction of pattern mixture specifications that avoid dependence on future responses.

### 3.3 Imputation methods

Imputation methods implicitly 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

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

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

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

#### 3.3.2 Last residual carried forward

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

$$Y_2 = \begin{cases} Y_{2a} & \text{if } R = 1 \\ \bar{Y}_2^C + (Y_1 - \bar{Y}_1) & \text{if } R = 0. \end{cases}$$

The implicit estimator is then

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

(4)

which is consistent for $\mu_{2a} + E(Z_{2a}|R = 1) - (1 - \pi)E(Z_1|R = 1)$. 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 $Z_1, Z_{2a}$ would then have the same sign. The expectation of $\bar{Y}_2$ will be closer to $\mu_{2a}$ than the expectation of $\bar{Y}_{2a}$, a desirable shift from the complete case estimand if $\mu_{2a}$ is the target for inference.

For these reasons last residual carried forward (LRCF) must be preferable to LOCF as a means of overcoming potentially informative drop-out, but in our opinion 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 this may also have been true at time 2, and consequently adjust the observed time 2 average accordingly. Second, this adjustment is down-weighted 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 $\bar{Y}_{2a}^C - (\bar{Y}_1^C - \bar{Y}_1)$ as another alternative.
3.3.3 Multiple imputation

One of several possible criticisms of both LOCF and LRCF 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 so as to enable estimation of, and if necessary adjustment for, the component of variation 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 missing at random (MAR) assumption is useful because it allows one part of the joint distribution to remain unspecified. MAR assumes that drop-out probability does not depend upon 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 \(C\) group is\[ L_{R|Y} = \prod_\mathcal{C}[R|Y_1] \quad L_Y = \prod \mathcal{C}[Y_1, Y_{2a}] \prod \mathcal{D}[Y_1]. \]The factorisation \([Y, R] = [R|Y][Y]\) is usually called a selection model (e.g. Michiels et al., 1999), although we prefer the term selection factorisation, to contrast with the pattern mixture factorisation \([Y, R] = [Y|R][R]\), and to emphasise 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

\[ V = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \] (5)

Then the maximum likelihood estimator of \(\mu_{2a}\) under MAR is\[ \hat{\mu}_{2a} = \bar{Y}^c_{2a} - \hat{\rho}(\bar{Y}^c_1 - \bar{Y}_1), \] (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 \(\bar{Y}^c_{2a} - (\bar{Y}^c_1 - \bar{Y}_1)\) as being appropriate when within-subject variability is small (\(\rho \to 1\)).

Parametric modelling under the combined assumption of MAR and separate parameterisation 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 the case, 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 administrative 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 \( \mu_{2a} \), or more generally covariate effects on the mean, then an alternative approach, still within the MAR framework, is to model \( \pi(Y_1) \) but leave \([Y_1, Y_{2a}]\) unspecified.

Under MAR we can estimate the drop-out probability 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 drop-out probability \( \hat{\pi}(Y_1) \), often via a logistic model. The marginal mean of \( Y_{2a} \) can now be estimated consistently 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):

\[
\hat{\mu}_{2a} = \frac{\sum_c \left( \frac{Y_{2a}}{1 - \hat{\pi}(Y_1)} \right)}{\sum_c \left( \frac{1}{1 - \hat{\pi}(Y_1)} \right)}.
\]

(7)

Use of (7) requires \( 1 - \hat{\pi}(Y_1) \) to be strictly positive for all subjects, and encounters difficulties in practice if this probability can be close to zero. This will not often be a material restriction within the current simplified setting, but 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 factorisation. In the special case of (2), the Diggle and Kenward model reduces to \((Z_1, Z_2)' \sim \mathcal{N}(0, \sigma^2V)\) with \( V \) as in (5), and

\[
\pi(Y_1, Y_2) = \frac{\exp(\alpha + \gamma_1 Y_1 + \gamma_0 Y_2)}{1 + \exp(\alpha + \gamma_1 Y_1 + \gamma_0 Y_2)},
\]

(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 is a parametrically testable special case of an MNAR model. Although the likelihood does not separate in the same way as under parametric MAR, likelihood inference is still possible by replacing \( \pi \) with its conditional expectation, derived from the conditional distribution of \( Y_2 \) given \( Y_1 \). The price paid for this facility is that correct inference now depends on two untestable modelling assumptions, the Normal model for \((Y_1, Y_2)\) and the logistic model for drop-out (Kenward, 1998). There is no closed form for the estimator of \( \mu_{2a} \).

### 3.7 Missing not at random: random effects

Under the Diggle and Kenward model the drop-out probability is directly determined by the responses \( Y_1 \) and \( Y_2 \), again assuming \( Y_{2a} = Y_{2b} \). 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

\[ Y_1 = \mu_1 + U + \epsilon_1 \quad Y_2 = \mu_2 + U + \epsilon_2 \]
\[ U \sim N(0, \tau^2) \quad \epsilon_1, \epsilon_2 \sim N(0, \sigma^2) \]
\[ \pi(U, \epsilon_1, \epsilon_2) = \pi(U) = \frac{\exp(\alpha + \gamma U)}{1 + \exp(\alpha + \gamma U)} \]

with independence between \( U, \epsilon_1 \) and \( \epsilon_2 \). Models of this type are in general 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 had the subject 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 EM or MCMC used, or the marginal likelihood can be obtained by numerical integration over \( U \), and the resulting likelihood maximised directly. Implementation is computationally intensive, even for this simple example, and there is again no closed form for \( \hat{\mu}_2 \).

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 paper to Robins et al. (1995), argue strongly for a more robust approach, on the assumption that the targets for inference involve only mean parameters. They again leave the joint distribution of responses unspecified but now model the drop-out probability as a function of both \( Y_1 \) and \( Y_{2a} \), for example by the logistic model (8). As applied within the simple framework of (2), the most straightforward version of the Rotnitzky et al. procedure 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 \( \alpha, \gamma_0 \) and \( \gamma_1 \) in (8) might be estimated by solving

\[ \sum_c \left\{ \frac{\hat{\pi}(Y_1, Y_{2a})}{1 - \hat{\pi}(Y_1, Y_{2a})} \right\} \phi(Y_1) - \sum_d \phi(Y_1) = 0, \]

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 \( \phi(Y_1) = (1, Y_1, Y_2)^T \). Since we only need \( \pi(Y_1, Y_{2a}) \) in the fully-observed group, all components of (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. Re-writing (9) as

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

\[
\hat{\mu}_{2a} = \frac{\sum_c \left\{ \frac{Y_{2a}}{1 - \hat{\pi}(Y_1, Y_{2a})} \right\}}{\sum_c \left\{ \frac{1}{1 - \hat{\pi}(Y_1, Y_{2a})} \right\}}.
\]

Rotnitzky et al. indicate that efficiency can be improved by augmenting the estimating equation for \( \mu_{2a} \) by a version of (9) (with a different \( \phi \)) and simultaneously solving both equations for all parameters. Fixed weight functions may also be introduced as usual. They also argue that estimation of the informative drop-out parameter \( \gamma_0 \) 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 \( \gamma_0 \) be treated as a known constant, but then varied over a range of plausible values so as to assess sensitivity of inferences for other parameters to the assumed value of \( \gamma_0 \).

Carpenter et al. (2006) compare inverse probability weighting (IPW) methods with multiple imputation. In particular, they consider a doubly robust version of IPW, 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 mis-specified. Their results show that doubly robust IPW out-performs the simpler version of IPW when the model for \( R \) is mis-specified, and out-performs multiple imputation when the model for \( Y_{2a} \) is mis-specified.

### 3.9 Sensitivity analysis

Rotnitzky et al. (1998) are not the only authors to suggest sensitivity analysis in this context. Other contributions include Copas and Li (1997), Scharfstein et al. (1999), Kenward (1998), Rotnitzky et al. (2001), Verbeke et al. (2001), Scharfstein et al. (2003), 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 \( \gamma_0 \).

### 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 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 formalises 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 common in event-history analysis; see, for example, Andersen et al. (1992). We do not claim that the proposed model 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

Longitudinal model

We suppose that $\tau$ measurements are planned on each of $n$ independent subjects. The measurements are to be balanced: that is, 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 $\tau$ 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 prior to each of the $\tau$ observation times. These we label $X_a$, noting that in theory there are also counterfactual covariates $X_b$: the values of covariates had a subject dropped out. We understand $X_a$ to be an $n \times p$ matrix process, constant if only baseline covariates are to be used, but potentially time-varying and possibly even dependent on the history of a subject or subjects. Note that we will 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 will follow this same convention for other processes.

At each observation time $t$ we acknowledge that the underlying hypothetical response may be measured with mean zero error $\epsilon_a(t)$. We assume that this process is independent of all others, and has the property that $\epsilon_a(s)$ and $\epsilon_a(t)$ are independent unless $s = t$. We make no further assumptions about this error process, and in particular 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 $\epsilon_a$, up to and including time $t$, by

$$G_t = \{X_a(s), Y_a(s), \epsilon_a(s) : s = 1, \ldots, t\}.$$}

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

$$G_{t-} = G_{t-1} \cup \{X_a(t)\},$$

12
which can be thought of as all information pertaining to \( X_a, Y_a \) and \( \epsilon_a \) 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 \( X_a(0) = Y_a(0) = \epsilon_a(0) = 0 \) 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

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

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 generalisation 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

\[
E\{\Delta Y_a(t)|\mathcal{G}_{t-}\} = X_a(t)\beta(t) - \epsilon_a(t - 1) \tag{10}
\]

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 \( E\{\Delta Y_a(t)|\mathcal{G}_{t-}\} \). 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 \( \beta \) the deterministic \( p \)-vector of regression functions representing the effects on the local characteristics of the covariates \( X_a \). Recall once again that \( \beta \) represents the hypothetical effects of covariates, assuming drop-out does not occur. Since \( \beta \) is an unspecified function of time, (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 other authors: see for example Lin and Ying (2001, 2003) or Martinussen and Scheike (2000, 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 accomodate measurement error by noting that, prior to time \( t \), no information is available about \( \epsilon_a(t) \), so the expected change in measurement error is simply \( -\epsilon_a(t - 1) \), 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 \( \sum_{s=1}^{t} \beta(s) \), with \( B(0) = 0 \), the transform of \( B \) by \( X_a \), denoted \( X_a \cdot B \), is given by

\[
(X_a \cdot B)(t) = \sum_{s=1}^{t} X_a(s) \{B(s) - B(s - 1)\} = \sum_{s=1}^{t} X_a(s) \beta(s)
\]

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 \( \beta \), both of which may vary over time.
The residual process is \( M_a = Y_a - X_a \cdot B - \epsilon_a \). 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 \( s \leq t \) we have that \( \mathbb{E}\{M_a(t)|\mathcal{G}_s\} = M_a(s) \), 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.

While 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, 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

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

The sample vector of martingale random effects is free to be, among other things, heteroskedastic, where the variance of a martingale may change over time and between subjects, and completely nonparametric, since the distribution of a martingale need not be specified by a finite dimensional parameter. We reiterate, however, that martingale residuals do 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, for example 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.

**Drop-out model**

Unfortunately, not all of the hypothetical longitudinal responses \( Y_a \) are observed. Rather, subject \( i \) gives rise to \( 1 \leq T_i \leq \tau \) measurements; that is, we observe \( Y_{ai}(1), \ldots, Y_{ai}(T_i) \). While both the hypothetical responses \( Y_{ai}(T_i+1), \ldots, Y_{ai}(\tau) \) and the realised 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 \( R_{it} \) denote an indicator process associated with subject \( i \), with \( R_{it}(t) = 1 \) if subject \( i \) is still under observation at time \( t \), and \( R_{it}(t) = 0 \) 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 in order 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 $R_{t-}$ the information available about drop-out prior to time $t$, and assume that $R_i(t) \in R_{t-}$. Although in this instance it follows that $R_{t-} = R_t$, it is useful to distinguish notationally between information available at these different points in time. We think of $R_t$ as a process in continuous time, but in practice 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, unrealised, event. Note that this does not preclude the possibility that future events might depend on past drop-out.

The second important requirement 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 $R_{t-}$ is the history of the drop-out process prior to time $t$. Censoring (or drop-out) is said to be independent of the hypothetical response processes $Y_a$ if, and only if

$$E\{\Delta Y_a(t)|G_{t-}, R_{t-}\} = E\{\Delta Y_a(t)|G_{t-}\}$$

(Andersen et al., 1992, p. 139). Independent censoring 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, had drop-out not occurred. This requirement is similar in spirit to the sequential version of MAR (S-MAR, Hogan et al., 2004, after Robins et al., 1995), which states that

$$[Y_a(t)|Y_a(s) : s < t; X_a(s), R(s) : s \leq t] = [Y_a(t)|Y_a(s) : s < t; X_a(s) : s \leq t].$$

We emphasise that independent censoring is a weaker assumption than S-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, while S-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, for example 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.

**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 $\beta$ in the local characteristics model,

$$E\{\Delta Y_a(t)|G_{t-}\} = X_a(t)\beta(t) - \epsilon_a(t - 1),$$

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

Recall that $R_i$ is an indicator process, unity if subject $i$ is still under observation. We shall
The key point is that the same parameters \( \beta \) are well-defined. So on our estimate the independence censoring assumption. The following model is induced for the observed longitudinal responses \( Y \):

\[
\mathbb{E}\{\Delta Y(t)|\mathcal{F}_{t-}\} = X(t)\beta(t) - \mathbb{E}\{\epsilon(t-1)|\mathcal{F}_{t-}\}
\]

(11)

where \( \epsilon = R \cdot \epsilon_a \). 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 \( \beta \) 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

Estimation

In order to estimate \( \beta = (\beta_1, \ldots, \beta_p)' \) we seek a matrix-valued process \( X^- \) having the property that \( X^- X = I \). However, due 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 endpoint 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 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 $B$ given by

$$\hat{B}(t) = \sum_{s=1}^{t} \hat{\beta}(s) = \sum_{s=1}^{t} X^{-}(s)\{Y(s) - Y(s-1)\} = (X^{-} \cdot Y)(t).$$  \hspace{1cm} (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 $B_{\mathcal{T}} = 1_{\mathcal{T}} \cdot B$, 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 $1_{\mathcal{T}}(t)\beta(t)$:

$$E\{\hat{\beta}(t) - 1_{\mathcal{T}}(t)\beta(t)\} = E\{X^{-}(t)\Delta Y(t) - 1_{\mathcal{T}}(t)\beta(t)\}$$
$$= E[X^{-}(t)E\{\Delta Y(t)|\mathcal{F}_{t-}\} - 1_{\mathcal{T}}(t)\beta(t)]$$
$$= E\{X^{-}(t)\{X(t)\beta(t) - E\{\epsilon(t-1)|\mathcal{F}_{t-}\}\} - 1_{\mathcal{T}}(t)\beta(t)\}$$
$$= E\{1_{\mathcal{T}}(t)\beta(t)\} - E\{\epsilon(t-1)\} - E\{1_{\mathcal{T}}(t)\beta(t)\} = 0$$

Therefore, $\hat{B}$ is unbiased for $B_{\mathcal{T}}$. What we have done is to mimic Aalen’s unbiased estimator, and 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 $\beta$ 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 zero step in their random effects; non-zero expected increments in the random effects cannot be distinguished from a change in population mean.

**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 $\epsilon$ 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 ordinary least squares regression, as outlined in the Appendix. In the application to follow, we use the bootstrap distribution for $\hat{B}$. 

17
4.3 Diagnostics

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

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

where $H = XX^-$ is the hat matrix of ordinary least squares and $Y^\gamma = 1_\gamma \cdot Y$. Standard residual plots, for example of $\hat{Z}$ against fitted least squares and $Y^\gamma = 1_\gamma \cdot Y$. Standard residual plots, for example of $\hat{Z}$ against fitted values or covariates, should reveal systematic mis-specifications 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 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 $t > 1$,

$$\text{Cov}\{M_a(1) + \epsilon_a(1), M_a(t) + \epsilon_a(t)\} = \text{V}\{M_a(1)\} = \text{V}\{M_a(s)\}$$

for all $1 \leq s < t$, 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 assumed by the process $\hat{Z}$, we have in particular that

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

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

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

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 Summary

In summary, our model is

\[ Y_a(t) = (X_a \cdot B)(t) + M_a(t) + \epsilon_a(t) \]

for \( t = 1, \ldots, \tau \). The observed data are \( R, X = RX_a \) and \( Y = R \cdot Y_a \). We assume that

\[ \mathbb{E}\{M_a(t)|X_a(t), R(t); X_a(s), R(s), Y_a(s), \epsilon_a(s) : s = 1, \ldots, t - 1\} = M_a(t - 1) \]

and our estimator for \( B \) is

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

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

5 Simple example revisited

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

\[ \mathbb{E}(Y_1) = \mu_1 \text{ and } \mathbb{E}(Y_{2a} - Y_1|Y_1, \epsilon_1) = \mu_{2a} - \mu_1 - \epsilon_1, \]

while the independent censoring assumption asserts that

\[ \mathbb{E}(Y_{2a} - Y_1|Y_1, \epsilon_1, R) = \mathbb{E}(Y_{2a} - Y_1|Y_1, \epsilon_1). \]

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

\[ Y_1 = \mu_1 + M_1 + \epsilon_1 \]  
\[ Y_{2a} = \mu_{2a} + M_{2a} + \epsilon_{2a} \]  
\[ \{(M_1, M_{2a}), \epsilon_1, \epsilon_{2a}\} \text{ mutually independent with zero means} \]  

and

\[ \mathbb{E}(M_{2a} - M_1|M_1, R = 1) = 0. \]

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

\[ \hat{\mu}_{2a} = \bar{Y}_1 + (\bar{Y}_{2a}^C - \bar{Y}_1^C) \]
\[ = \bar{Y}_{2a}^C - (\bar{Y}_1^C - \bar{Y}_1) \]

and is unbiased for \( \mu_{2a} \).

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 (17) are needed to give \( \mu_1 \) and \( \mu_{2a} \) interpretations as drop-out-free population means, the parameters of interest. Note, though, that we do not require \( M_1, M_{2a} \) to be independent. Equation (18) provides our key assumption, that the subject-specific random effects have mean zero 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 (18), taken together with (17), is that the subject-specific random effects also have mean zero 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: (18) holds if, and only if

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

where \( \Delta = M_{2a} - M_1 \). Examples that satisfy the above 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 \( \Delta \) for which \( E(\Delta|M_1) = 0 \); any \( \pi(M_1, \Delta) \) that is an even function of \( \Delta \), taken together with any zero mean, symmetric distribution \( [\Delta|M_1] \).

None of the these examples are 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 S-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 (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.

Scenario 1

The first simulation scenario mimics the schizophrenia example 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 zero 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

\[
S_a(t) = U_0 + U_1 w_t \quad M_a(t) = U_0 + U_1 + U_2 + \cdots + U_{t-1}.
\]

Then \( S_a \) is a random intercept and slope process, of the kind described by Laird and Ware (1982), while \( M_a \) is a martingale. We take \( \mathbf{V}(U_0) = \sigma_0^2 I \), \( \mathbf{V}(U_1) = \sigma_1^2 I \) and choose the variances of the further values to ensure that \( \mathbf{V}\{S_a(t)\} = \mathbf{V}\{M_a(t)\} \). This setup 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

\[
Y^{S}_a(t) = \mu_t + S_a(t) + \epsilon_a(t) \quad Y^{M}_a(t) = \mu_t + M_a(t) + \epsilon_a(t)
\]
with \( \epsilon_a(t) \sim N(0, \sigma_a^2 I) \), 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 \( n = 125, 250, 500, \) and \( 1000 \) we took 1000 simulations from this model. We used \( \mu_1 = \cdots = \mu_6 = 0 \) and chose the other parameter values to correspond roughly to the schizophrenia data: \( \sigma_0^2 = 200, \sigma_1^2 = 15, \sigma_\epsilon^2 = 100, \) and

\[
(\alpha_1, \ldots, \alpha_5) = (-8, -6, -6, -6, -4) \quad (\gamma_1, \ldots, \gamma_3) = (0.2, 0.3, 0.3, 0.5, 0.6).
\]

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 using our linear increments (LI) approach, an inverse-probability weighted estimating equation approach (IPW), and by fitting a multivariate Normal distribution with unstructured within-subject covariance matrix (UMN). Under both IPW and UMN we made a mis-specified 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 MAR.

Table 1 summarises results at \( n = 500 \). There was severe downward bias in the observed mean values (OLS) for each of \( Y^M_a \) and \( Y^S_a \) and this is only partly corrected by the mis-specified IPW or UMN. 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 the model is then also mis-specified. Usually such mis-specification would be detected by the diagnostics. For example, boxplots of the residual covariances (Figure 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).

**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

\[
B(t) = (0, 1(t > 0) e^{-t(t-1)}, t)'.
\]

We add to the mix some error in measurement \( \epsilon \), arising according to a \( t \)-distribution on \( \nu \) degrees of freedom and scaled by a factor \( \sigma_\epsilon \); that is, \( \sigma_\epsilon^{-1} \epsilon(t) \sim t(\nu) \). The final measurement times \( T_1, \ldots, T_n \) are determined by the relationship

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

so that \( 1 \leq T_i \leq 7 \) for each \( i \).

We defined

\[
Y^S_a = X_a \cdot B + S_a + \epsilon_a \quad Y^M_a = X_a \cdot B + M_a + \epsilon_a.
\]

The parameters were taken to be

\[
\sigma_0 = \sigma_1 = \sigma_W = 1, \pi = 1/2, \sigma_\epsilon = 1/3, \nu = 3, \alpha = -7.
\]

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).
Table 1: Estimated mean responses and standard errors (SE) for Scenario 1 using observed data without correction for drop-out (OLS), with inverse probability weighting (IPW) or a multivariate Normal model with unstructured covariance matrix (UMN), both of which falsely assume MAR, and under the linear increments (LI) method. Coverage (Cov) of nominal 95% confidence intervals under LI also included. Sample size was $n = 500$, and results were averaged over 1000 simulations.

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

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

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$n = 125$</th>
<th>250</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Power</td>
<td>0.307</td>
<td>0.530</td>
<td>0.766</td>
</tr>
<tr>
<td></td>
<td>Type I Error</td>
<td>0.056</td>
<td>0.056</td>
<td>0.053</td>
</tr>
<tr>
<td>2</td>
<td>Power</td>
<td>0.147</td>
<td>0.241</td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>Type I Error</td>
<td>0.056</td>
<td>0.059</td>
<td>0.045</td>
</tr>
</tbody>
</table>
Figure 1: Boxplots of $\text{Cov}(\hat{Z}(1), \hat{Z}(t))$ based on 1000 simulations under Scenario 1 at sample size $n = 500$. Left plot, true martingale structure $Y^M$; right plot, Laird-Ware random intercept and slope structure $Y^S$. 
Figure 2: Summary of estimates $\hat{B}$ for Scenario 2, at sample size $n = 500$. Mean dynamic linear estimates from $Y^M$ (solid lines) and $Y^S$ (dashed lines) together with true values (dotted lines).

Mean estimates of $B$ for sample size $n = 500$ using both $Y^M$ and $Y^S$ are shown in Figure 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 the mis-specified 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 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, amongst whom 66 dropped out for reasons unrelated to their underlying condition. The remaining 183 represent potentially informative drop-out, though we emphasise that our new approach does not need to distinguish these from the non-informative drop-outs. We mention them only because we will 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 baseline ($t = 1$) and thereafter at weeks 1, 2, 4, 6 and 8 ($t = 2, 3, 4, 5$ and 6) of the study. The only covariates used here are treatment groups. The dotted lines in Figure 3 show for reference the observed
Figure 3: Estimated PANSS mean values under ordinary least squares (dotted lines) and our dynamic linear approach (solid lines). The topmost lines correspond to placebo, the middle lines to standard treatment, and the lowest lines to experimental treatment.

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.

Figure 3 displays the pronounced differences between the ordinary least squares (OLS) estimates and their dynamic linear counterparts. The OLS estimates invite the counterintuitive 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 placebo do not improve over time; in fact, there is even a suggestion that their PANSS scores increase slightly. The leveling out of treatment effects over time seen under our new approach is also unsurprising.

In Figure 4 and Table 3 we compare the dynamic linear fits with those obtained under four other approaches. Figure 4 shows the estimated means for each treatment group while 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

A. Maximum likelihood estimation under a multivariate Normal model with unstructured
covariance matrix (UMN). This approach assumes MAR.

B. A quadratic random effects joint longitudinal and event-time informative drop-out model fitted by Dobson and Henderson (2003) using EM estimation, as suggested by Wulfsohn and Tsiatis (1997). Dobson and Henderson (D+H) compared four random effects structures and concluded that, between these, the model used here with random intercept, slope and quadratic terms “is strongly preferred by likelihood criteria, even after penalizing for complexity”.

C. An inverse probability weighted (IPW) estimating approach as described by Robins et al. (1995), with a logistic MAR model for drop-out.

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 for event history analyses desibed 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 MAR models. Method B assumes a Gaussian response but method C has no modelling assumptions for the responses, a gain obtained at the expense of an increase in standard errors. Method D leads to estimates that are comparable to the fit obtained using only exogenous covariates, albeit slightly closer to the observed means. Method B, the quadratic random effects model, gives estimates close to those obtained 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 needed to compute the standard errors were completed in under 10 seconds on an unremarkable laptop computer. In an appendix, we demonstrate briefly one way in which our dynamic linear models may be implemented 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 are necessary under our proposed approach. Given the similarities between our dynamic linear results and those of B, the Dobson and Henderson assumption that these other events are uninformative about PANSS seems to be justified.

The proposed diagnostics may be illustrated using these data. Having computed $\hat{B}$, it is straightforward to extract $\hat{Z}$. Figure 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. The panel for week 1 clearly indicates a weak negative association, 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. Panel D of Figure 4 shows that the fitted mean response profiles are not materially affected by the misspecification indicated by Figure 5.

Boxplots illustrating the bootstrap distribution of the diagnostic $n^{-1}\hat{Z}'(1)\hat{Z}(t)$ are shown in Figure 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 boxplot may be safely ignored. Based on the remaining boxplots, 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,
Figure 4: Estimated PANSS mean values for (from top to bottom, in every case) placebo, standard and experimental treatment groups. The dashed lines correspond to the estimates generated under methods A–D described in the text. Common to all four plots are the solid lines, the estimates under the dynamic linear approach.

Table 3: Effect of treatment on change in mean response (week eight minus week zero) under the linear increments (LI) approach (12), ordinary least squares with an independence assumption (OLS) and methods A–D described in the text. ‘S’ represents the standard treatment, ‘P’ placebo, and ‘E’ the experimental treatment. Standard errors are in parentheses.
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.

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 $Y_{\text{obs}}$ and missingness indicators $R$. We have argued that this set-up can be too simple, as it does not recognise 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 well 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 as to 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 some 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 well 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 \textit{all} future values to be independent of the drop-out decision: the realised response is free to depend on this decision. Neither 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 MAR form. Ultimately, however, both MAR and the independent censoring assumption share the same purpose: to enable inference by making assumptions about the dropout process. 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, Figure 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) advocate 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 dropout. 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’, in order to draw inference about the longitudinal features a population would have exhibited, assuming 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.
Appendix. Fitting dynamic linear models 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 using the `lmList` command from the `nlme` package (Pinheiro and Bates, 2000) in R or S-PLUS. For example, in order 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 associated with the given subject and time-point: that is, it contains $\Delta Y_i(1), \ldots, \Delta Y_i(T_i)$ for every $i$. Then

```r
> 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

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

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

```r
> 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

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

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

ACKNOWLEDGEMENTS

The authors are grateful for the detailed comments and helpful advice of all referees for the paper. Peter Diggle is supported by an EPSRC Senior Fellowship. Daniel Farewell’s research was carried out during his MRC-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.

REFERENCES


