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,
Neuadd Meirionnydd, Heath Park, Cardiff, CF14 4YS.
Email: farewell1d@cf.ac.uk

December 15, 2006

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 valid inference for the underlying drop-out-free population is possible provided one important assumption holds: that 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 does not depend on 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), Shi (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 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 individuals exhibits a certain stability prior to drop-out. Thirdly, such stability is maintained beyond each drop-out time by the diminishing sub-set 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 \texttt{S} language. Appendix B gives proofs of two propositions.

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. For recent reviews of joint modelling, see for example Hogan \textit{et al.} (2004) or Tsiatis and Davidian (2004). Certainly the scientific goals of joint modelling can differ from ours: a common scenario sees the time-to-event as a survival time and the longitudinal measurements as time-varying covariates that are measured with error, a scenario we discuss no further. However, longitudinal data with drop-out can formally be considered as joint modelling in which the time-to-event is the drop-out time; see, for instance, Henderson \textit{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.

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 $\mu_t = \mathbf{E}(Y_t)$, 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 $n$ 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:

$$Y_1 = \mu_1 + Z_1 \quad Y_2 = \mu_2 + Z_2 \quad \mathbf{E}(Z_1) = \mathbf{E}(Z_2) = 0.$$  (1)

The parameter $\mu_1$ is the population mean at time 1. Writing down (1) invites a similar inter-
pretation 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

$$Y_1 = \mu_1 + Z_1 \quad Y_{2a} = \mu_{2a} + Z_{2a} \quad Y_{2b} = \mu_{2b} + Z_{2b} \quad \Pr(R = 0|\mathcal{S}) = \pi(\mathcal{S}). \quad (2)$$

In (2), $\mathbb{E}(Z_1) = \mathbb{E}(Z_{2a}) = \mathbb{E}(Z_{2b}) = 0$, $\mathcal{S}$ denotes a set of conditioning variables and we allow $\pi(\cdot)$ to depend arbitrarily on $\mathcal{S}$. We make no assumption of independence between $Z_1, Z_{2a}, Z_{2b}$, and for the unconditional case $\mathcal{S} = \emptyset$ we write $\pi = \pi(\emptyset) = \Pr(R = 0)$. By construction, the parameters $\mu_1, \mu_{2a}$ and $\mu_{2b}$ are the marginal expectations of $Y_1, Y_{2a}$ and $Y_{2b}$ respectively.

Now, we collect some data. We have $n$ independent subjects and for each of these we observe the values of $Y_1$ and $R$. Additionally, we observe $Y_{2a}$ for subjects who have $R = 1$, and who will form the group $C$. Unfortunately $Y_{2a}$ is not observed for those with $R = 0$, the group $D$. At least equally unfortunately, $Y_{2b}$ is not observed for any subject.

Returning to the context of longitudinal data with drop-outs, the groups $C$ and $D$ are the completers and drop-outs, while $Y_1$ and $R$ have the obvious interpretations as the response at time 1 and the drop-out indicator, respectively. If $R = 1$ then $Y_{2a}$ is the longitudinal response at time 2, whereas if $R = 0$ it is the counterfactual that would have been observed had the subject not dropped out. If subjects do drop out then $Y_{2b}$ is the extant but unobserved longitudinal response at time 2; if they do not drop out it is another counterfactual, what would have occurred if the subject in question had 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 $\mu_1 = \mathbb{E}(Y_1)$ 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 $\mathbb{E}(Y_2)$. 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 happens to be unobserved for subjects in group \( D \). This would be appropriate if we think that the problem of dealing with drop-out would be solved if observation could have been continued after the event \( R \) is determined; depending on the precise reason for drop-out, continued observation is sometimes possible, for example when a subject withdraws from a trial but remains under observation.

Because the definition (3) means that \( Y_2 \) is unobserved in group \( D \), further progress will depend on the strong and untestable assumption that \( Y_{2a} \equiv Y_{2b} \). 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.

The data we analyse in Section 7 come from a longitudinal randomised clinical trial of drug treatments for schizophrenia; the response of interest was a measure of psychiatric disorder, recorded on the Positive And Negative Symptom Scale (PANSS) and scheduled to be measured on six occasions. In this application, drop-out implies discontinuation of the assigned drug. Objective 1 would apply if the scientific aim was an intention-to-treat analysis, as opposed to a per-protocol analysis.

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 individuals were removed from the study because their condition did not improve. The second objective would therefore be appropriate in this context if interest were in the PANSS score recorded for individuals 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 \( D \) individuals, is

\[ Y_2 := Y_{2a}, \]
which is appropriate if scientific interest lies in the (possibly hypothetical) per-protocol response distribution of a drop-out-free population, one in which $\pi = 0$. The assumption $Y_{2a} \equiv Y_{2b}$ makes Objectives 1 and 3 equivalent.

Note that we do not interpret $Y_{2a}$ as for Objective 2, the response at time 2 conditional on $R = 1$. The potential difference between marginal and conditional means makes this clear. To illustrate, suppose that $Z_{2a}$ is $t$-distributed, and that

$$
\pi(Z_1, Z_{2a}, Z_{2b}) = \pi(Z_{2a}) = \begin{cases} 
1 & Z_{2a} \leq 0 \\
0 & Z_{2a} > 0 
\end{cases}
$$

so that there is 50% drop-out in large samples. The difference $\Delta$ between the unconditional expectation of $Y_{2a}$ and its conditional expectation given $R = 1$ is shown below, along with the conditional standard deviation (CSD) of the observed (positive) $Z_{2a}$, when $Z_{2a}$ has $r$ degrees of freedom.

<table>
<thead>
<tr>
<th>$r$</th>
<th>$\infty$</th>
<th>10</th>
<th>5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$</td>
<td>0.80</td>
<td>0.86</td>
<td>0.95</td>
<td>1.10</td>
</tr>
<tr>
<td>CSD</td>
<td>0.60</td>
<td>0.71</td>
<td>0.88</td>
<td>1.34</td>
</tr>
</tbody>
</table>

Clearly, $\Delta$ can be relatively large. Assuming that subjects with negative $Z_{2a}$ always drop out might seem extreme but we need to recall that this would never be identified from the observed data, as a selection effect is indistinguishable from a change in the mean response. It is easy to construct other illustrations where the unconditional and conditional means are even more substantially different.

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 individual 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, Dufoil 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.

We note that the hypothetical target for inference, Objective 3, is analogous to the usual estimand in event-time analysis, with drop-out equivalent to censoring. There, the observed event intensity $\lambda(t)$ is factorised as

$$
\lambda(t) = R(t)\alpha(t)
$$

separating an at-risk indicator $R(t)$ from the hypothetical event intensity $\alpha(t)$. This latter is our inferential objective, a deterministic function referring to a hypothetical population in
which individuals never drop out and are always at risk of an event. In the longitudinal setting, \( \mu_{2a} \) is the mean of a hypothetical population in just the same way.

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 sub-groups 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

We now concentrate on Objective 3, sometimes adding the assumption that \(Y_{2a} \equiv Y_{2b}\) so that 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 $\bar{Y}_2^C (1 - \hat{\pi}) + Y_1^D \hat{\pi}$, where $\bar{Y}_1^D$ is the mean at time 1 for group $D$. The estimator is consistent for

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

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

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

The implicit estimator is then

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

which is consistent for $\mu_{2a} + \mathbb{E}(Z_{2a} | R = 1) - (1 - \pi) \mathbb{E}(Z_1 | R = 1)$. 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 $\bar{Y}_{2a}^C$, 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.
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 \([Y_1, Y_{2a}]\).

The likelihood contribution in the \(C\) group is
\[
P(R = 1|Y_1, Y_{2a}) \prod Y_{2a} Y_1 = (1 - \pi(Y_1)) Y_1 Y_{2a},
\]
whilst in the \(D\) group it is just \(\pi(Y_1)|Y_1\). The combined likelihood is thus
\[
L = L_{R|Y} \times L_Y,
\]
where
\[
L_{R|Y} = \prod R|Y_1, L_Y = \prod [Y_1, Y_{2a}] Y_1.
\]

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

Then the maximum likelihood estimator of \(\mu_{2a}\) under MAR is
\[
\hat{\mu}_{2a} = \bar{Y}_{2a}^C - \hat{\rho}(\bar{Y}_1^C - \bar{Y}_1),
\]
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 \(\bar{Y}_{2a}^C - (\bar{Y}_1^C - \bar{Y}_1)\) as being appropriate when within-individual 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, provided that the target for inference is the per-protocol response distribution. 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 Horwitz-Thompson weighted average of the observed $Y_{2a}$, where the weights are the inverse probabilities of observation (Robins et al., 1995):

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

(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 sub-groups of subjects.

Carpenter and Kenward (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.

3.6 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 $\mathbf{Z}_1, \mathbf{Z}_2 \sim N(0, \sigma^2 V)$ 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 there are now two untestable modelling assumptions: the Normal model for $(Y_1, Y_2)$ and the logistic model for drop-out. 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, and has certainly been more widely studied. 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
\[
\begin{align*}
Y_1 &= \mu_1 + U + \epsilon_1 \\
U &\sim N(0, \tau^2) \\
\epsilon_1, \epsilon_2 &\sim N(0, \sigma^2)
\end{align*}
\]
\[
\pi(U, \epsilon_1, \epsilon_2) = \pi(U) = \frac{\exp(\alpha + \gamma U)}{1 + \exp(\alpha + \gamma U)}
\]
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 \(\hat{\mu}_{2a}\).

Models of this kind are conceptually attractive, and parameters are identifiable without any further assumptions. But the inferences are also greatly dependent on the precise structure given to joint distributions, a structure that is 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 \(\phi(Y_{1}) = (1, Y_{1}, Y_{1}^2)'\). 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

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

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 \(\gamma_0\).

### 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. The widely-used MAR assumption, although often useful in practice as empirical models, may be philosophically unattractive in the setting of discrete-time observation of a continuous-time process. Specifically, except in very particular circumstances it is hard to imagine how the probability of drop-out at time \(t\) could depend on an observed measurement at time \(t - 1\) but not on unobserved measurements in the interval \((t - 1, t)\). Proposed MNAR models are often only weakly identified, and many authors have suggested a detailed sensitivity analysis as their preferred analysis strategy. In the next section we offer an alternative, which is related to existing work in event-history analysis and, by working with increments of the measurement process, focuses explicitly on the prediction of future from past behaviour.
4 Proposal

In this section we introduce a new approach to the analysis of longitudinal data with potentially response-dependent drop-out. Assumptions are presented in the language of stochastic processes, but these are restated more informally in Section 5 for our simple example, to bring out the intuition of the proposal. For background about stochastic processes and martingales see, for example, the excellent books by Andersen et al. (1992), Grimmett and Stirzaker (2001) and Neveu (1975).

4.1 Model specification

Longitudinal model

We suppose that $\tau$ measurements are planned on each of $n$ independent individuals. The measurements are to be balanced: that is, the intended observation times are identical for each individual, and without loss of generality we label these times $1, \ldots, \tau$. For the time being, let us suppose that all $n$ individuals do indeed provide $\tau$ measurements. In the notation of Section 2, $Y_a$ is therefore observed for every individual 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 an individual 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 an individual or individuals. 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

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

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

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

which can be thought of as all information pertaining to $X_a, Y_a$ and $\epsilon_a$ 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 individuals.

For convenience, we set $X_{at}(0) = Y_{at}(0) = \epsilon_{at}(0) = 0$ 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

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

(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 accommodate 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 $E\{M_a(t)|\mathcal{G}_s\} = M_a(s)$, 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 these 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

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

The sample vector of martingale random effects is free to be, among other things: differentially distributed, so that individuals need not all be drawn from some larger, hypothetical distribution; specifically, heteroskedastic, where the variance of a martingale may change over time, and between individuals; 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. But the very fact that, before examining the data, it may be assessed sufficiently to be dismissed, is one of the chief merits of the approach. Without the need for complicated diagnostics, applied scientists with no background in stochastic processes may assess, not the truth, but the plausibility of the martingale hypothesis. And it seems to us that in many applications an underlying martingale structure seems credible, at least as a first approximation. If necessary, non-linear models for the local characteristics are certainly possible, the famous Cox (1972) model being an obvious example, but we discuss these no further here. We reiterate that the linear model may be adapted to include summaries of previous longitudinal responses: the martingale hypothesis may sometimes be invalid when using only exogenous covariates, but appropriate when dynamic covariates are included, for example summaries of the individual trajectories to date.

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, individual $i$ 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 $R_i$ denote an indicator process associated with individual $i$, with $R_i(t) = 1$ if individual $i$ is still under observation at time $t$, and $R_i(t) = 0$ otherwise. We let $\mathcal{R}_i$ be the history of these indicator processes up to time $t$. Note that we do not distinguish between competing types of drop-out, for instance between
administrative censoring, treatment failure or death. This is not because we could not draw such a distinction, if this were of scientific interest, but because we need not in order to make inferences regarding the hypothetical responses $Y_a$, which refer to a population in which there is no drop-out of any kind.

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 available about drop-out prior to 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 available at these different points in time. We think of $R_i$ 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 $\mathcal{R}_{t-}$ 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

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

(Andersen et al., 1992, p. 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 individuals, had drop-out not occurred. This requirement is very 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 is a statement about conditional means, while S-MAR concerns conditional distributions. Also, S-MAR does not condition on past measurement error, so there is incomplete knowledge of $M_a$ and $\epsilon_a$. Nevertheless, the technical requirements that, firstly, the $R_i$ be predictable and, secondly, that censoring be independent, are reminiscent of MAR and of its many generalisations. Where MAR demands that the probability of drop-out be independent of unobserved data, we require that drop-out be independent of the (potentially unobserved) future. Ultimately, both MAR and the independent censoring assumption share the same purpose: MAR is a sufficient condition to draw inference using the observed data likelihood, while independent censoring is a sufficient condition to draw inference using the observed local characteristics.

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.
As we have already discussed, our target for inference will be the hypothetical effects of covariates supposing, contrary to fact, that individuals did not drop out of observation. More explicitly, we seek to make inference about \( E[f_Y a(t) | \mathcal{T}_t] \) in the local characteristics model, 

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

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 \( R_i(t) \) is an indicator process, unity if individual \( i \) is still under observation. We shall write 

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

for the diagonal matrix with the \( R_i(t) \) along the diagonal. We claim that the processes \( R, X = RX_a \) and \( Y = R \cdot Y_a \) 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 \( R \cdot Y_a \) is the transform of \( Y_a \) by \( R \), and is defined by 

\[
(R \cdot Y_a)(t) = \sum_{s=1}^{t} R(t) \Delta Y_a(t).
\]

So \( R \cdot Y_a \) is the process \( Y_a \) whose individual elements are stopped, that is, held constant, after the time \( T_i \) of their last observations. Hence this process, too, is observable. We denote the history of the observed data \( X, Y \) and \( R \) as 

\[
\mathcal{T}_t = \{X(s), Y(s), R(s) : s = 1, \ldots, t\},
\]

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

\[
E\{\Delta Y(t) | \mathcal{T}_{t-} \} = X(t) \beta(t) - E\{\epsilon(t - 1) | \mathcal{T}_{t-} \} \tag{11}
\]

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 both the local characteristics of \( Y \) 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**

We present a non-parametric approach to the estimation of \( \beta \) and, by extension, of \( B \). This decision is essentially pragmatic, for non-parametric estimates of \( B \) are available in closed form and consequently are readily understood. Yet there is a philosophical advantage, too: we no
In order to estimate $\beta = (\beta_1, \ldots, \beta_p)'$ we seek a process $X^-$ having the property that $X^-X = I$. However, due to drop-out such a matrix does not always exist. Let $\mathcal{T} = \{t : \text{det}\{X'(t)X(t)\} \neq 0\}$, 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 individuals 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

$$\hat{B}(t) = \sum_{s=1}^{t} \hat{\beta}(s) = \sum_{s=1}^{t} X^-(s)\{Y(s) - Y(s-1)\} = (X^- \cdot Y)(t). \quad (12)$$

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 $1_\mathcal{T}(t)\beta(t)$:

$$\mathbb{E}\{\hat{\beta}(t) - 1_\mathcal{T}(t)\beta(t)\} = \mathbb{E}\{X^-(t)\Delta Y(t) - 1_\mathcal{T}(t)\beta(t)\}$$

$$= \mathbb{E}\{X^- \mathbb{E}\{\Delta Y(t)|\mathcal{F}_{t^-}\} - 1_\mathcal{T}(t)\beta(t)\}$$

$$= \mathbb{E}\{X^- \mathbb{E}\{X(t)\beta(t) - \mathbb{E}\{\epsilon(t-1)|\mathcal{F}_{t^-}\}\} - 1_\mathcal{T}(t)\beta(t)\}$$

$$= \mathbb{E}\{1_\mathcal{T}(t)\beta(t)\} - \mathbb{E}\{\epsilon(t-1)\} - \mathbb{E}\{1_\mathcal{T}(t)\beta(t)\} = 0$$

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 individual 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**

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 $\epsilon_a$ 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 $B$. Though we provide a result that $\hat{B}$ is $\sqrt{n}$-consistent for $B$ with a Gaussian limiting distribution (Proposition 2), we use the bootstrap distribution for $\hat{B}$ in the application to follow.

For very large datasets, it can be useful to have an approximate variance estimate that is easily computable. If it can be assumed that measurement error is insubstantial, it is possible to exploit further the analogy with ordinary least squares and extract a naive estimate of the sampling variability of $\hat{B}$ without the need for bootstrap computations (Proposition 1). For simplicity of the approximation, we also assume that the variances of the martingale increments are deterministic, and do not depend on the past evolution of the process.

**Proposition 1** Suppose that $\epsilon_a(t) = 0$ for every $t$, so that measurement error is negligible. Further assume that $M_a$ is a vector of independent, identically distributed martingales satisfying $\mathbf{V}\{\Delta M_a(t) | \mathcal{F}_-\} = \sigma^2(t)I$, where $I$ is the $n \times n$ identity matrix. Then

$$\sum_{s=1}^{t} 1_{\mathcal{F}}(s)\{X'(s)X(s)\}^{-1}\hat{\sigma}^2(s)$$

is unbiased for $\mathbf{V}\{\hat{B}(t) - B_{\mathcal{F}}(t)\}$, where

$$\hat{\sigma}^2(t) = \frac{\Delta Y'(t)(I - XX^{-})\Delta Y(t)}{\text{tr } R(t) - r_X}.$$

Here $\text{tr } R(t)$ is the trace of $R(t)$ and $r_X$ is the rank of $X$.

The proof is given in Appendix B.

The variance estimator of Proposition 1 is just the cumulative sum of the variance estimates of individual least squares fits at each $t \in \mathcal{F}$. If the error-free assumption is not justified, the contribution of measurement error is absorbed into the martingale random effects. Because of its negatively correlated increments, this procedure will tend to overestimate the variance of $\hat{B}$.

We now make a statement about the asymptotic behaviour of $\hat{B}$ in the presence of measurement error.

**Proposition 2** Suppose that individuals are independent and identically distributed, and let

$$d(s, t) = \mathbf{E}\{X'_1(s)X_1(t)\} = \cdots = \mathbf{E}\{X'_n(s)X_n(t)\} \text{ for all } s, t$$

$$v = \mathbf{V}(X'_1 \cdot M) = \cdots = \mathbf{V}(X'_n \cdot M)$$

$$e = \mathbf{V}(\epsilon_1) = \cdots = \mathbf{V}(\epsilon_n),$$

with $d(t, t) = d(t)$ and $\det d(t) \neq 0$ for every $t$. Then, as $n \to \infty$, $\sqrt{n}(\hat{B} - B)$ converges (pointwise, in distribution) to a Normal random process with mean zero and variance process $(d^{-1}, d^{-1}) \cdot v + w$, where

$$\{(d^{-1}, d^{-1}) \cdot v\}(t) = \sum_{s=1}^{t} d^{-1}(s)\{v(s) - v(s - 1)\}d^{-1}(s)$$
and

\[ w(t) = d^{-1}(t)e(t) + \sum_{s=1}^{t-1} \{d^{-1}(s) - d^{-1}(s)d(s, s+1)d^{-1}(s+1) \}
- d^{-1}(s+1)d(s+1,s)d^{-1}(s) + d^{-1}(s+1)\}e(s). \]

In the case of constant covariates, this reduces to

\[ w(t) = 2d^{-1}(t)e(t) - (d^{-1} \cdot e)(t). \]

The proof of this proposition is unilluminating and therefore omitted, but details are available on request. Proposition 2 shows that \( \hat{B} \) is consistent for \( B \) and asymptotically normal, but also reveals why variance estimation is problematic. The contribution \( w \) of the measurement error is not \( d \cdot e \) as might be expected, but is a complicated function of the covariate structure. This expression is only slightly simplified if covariates do not depend on time.

### 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

\[ \hat{Z} = (I - H) \cdot Y^\beta, \]

where \( H = XX^- \) is the hat matrix of ordinary least squares. 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 individuals.

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 \( t > 1 \),

\[ \text{Cov}\{M_a(1) + \epsilon_a(1), M_a(t) + \epsilon_a(t)\} = \text{V}\{M_a(1)\} \]  \hspace{1cm} (13)

This result is easily proved, since martingales have uncorrelated increments and the errors \( \epsilon \) are mutually independent. The point about (13) is that although the left-hand-side may be evaluated at each of \( t = 2, 3, 4, \ldots \), the right-hand-side is a constant. Proposition 3 establishes that a plot of \( \text{Cov}\{\hat{Z}(1), \hat{Z}(t)\} \) against \( t \) has diagnostic value, with departures from a straight line with zero slope indicating unsuitability of (11).

**Proposition 3** For \( t = 2, \ldots, \tau \), \( E\{\hat{Z}(1)\hat{Z}(t)\} \) is a scalar constant.
The proof is given in Appendix B.

Clearly similar plots may be derived based on the observation that
\[
\text{Cov}\{M_a(s) + \epsilon_a(s), M_a(t) + \epsilon_a(t)\} = \text{V}\{M_a(s)\}
\]
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
\[
\mathbb{E}\{\hat{Z}'(1)\hat{Z}(2)\} = \mathbb{E}\{\hat{Z}'(1)\hat{Z}(\tau)\}.
\]
Therefore \(\mathbb{E}\{\hat{Z}'(1)(\hat{Z}(\tau) - \hat{Z}(2))\} = 0\), and, for large \(n\) the approximation
\[
\frac{\hat{Z}'(1)(\hat{Z}(\tau) - \hat{Z}(2))}{\sqrt{\text{V}[\hat{Z}'(1)(\hat{Z}(\tau) - \hat{Z}(2))]}} \sim N(0, 1)
\]
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.

## 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:
\[
\mathbb{E}(Y_1) = \mu_1 \quad \text{and} \quad \mathbb{E}(Y_{2a} - Y_1|Y_1, \epsilon_1) = \mu_{2a} - \mu_1 - \epsilon_1,
\]
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
\[
Y_1 = \mu_1 + M_1 + \epsilon_1 \quad \text{(15)}
\]
\[
Y_{2a} = \mu_{2a} + M_{2a} + \epsilon_{2a} \quad \text{(16)}
\]
\[
\{(M_1, M_{2a}), \epsilon_1, \epsilon_{2a}\} \text{ mutually independent with zero means} \quad \text{(17)}
\]
and
\[
\mathbb{E}(M_{2a} - M_1|M_1, R = 1) = 0 \quad \text{(18)}
\]
Under assumptions (15)–(18), our least squares estimator (12) is given by
\[
\hat{\mu}_{2a} = \hat{Y}_1 + (\hat{Y}_{2a} - \hat{Y}_1^c)
\]
\[
= \hat{Y}_{2a}^c - (\hat{Y}_1^c - \hat{Y}_1) \quad \text{(19)}
\]
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 \textit{conditional on staying in the study}. 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 \textit{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

1. a random intercept model in which \( \Delta \equiv 0 \), with any \( \pi(\cdot) \);
2. an independent censoring drop-out model in which \( \pi(M_1, \Delta) = \pi(M_1) \), with any \( \Delta \) for which \( E(\Delta|M_1) = 0 \);
3. 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 above 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 addition 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 $V(U_0) = \sigma_0^2 I$, $V(U_1) = \sigma_1^2 I$ and choose the variances of the further values to ensure that $V(S_a(t)) = V(M_a(t))$. 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

$$Y_a^S(t) = \mu_t + S_a(t) + \epsilon_a(t) \quad Y_a^M(t) = \mu_t + M_a(t) + \epsilon_a(t)$$

with $\epsilon_a(t) \sim N(0, \sigma_2^2 I)$, and independence between time-points. The drop-out probabilities at measurement times $t = 1, \ldots, 6$ 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: $\sigma_0^2 = 200$, $\sigma_1^2 = 15$, $\sigma_2^2 = 100$, and

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

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 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. There was severe downward bias in the observed mean values (Obs) for each of $Y_a^M$ and $Y_a^S$ 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

$$B(t) = (0, 1\{t > 0\} e^{-(t-1)}, t).$$

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
<table>
<thead>
<tr>
<th>$Y^M$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs</td>
<td>Mean</td>
<td>-0.00</td>
<td>-0.30</td>
<td>-2.75</td>
<td>-4.34</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.77</td>
<td>0.78</td>
<td>0.77</td>
<td>0.91</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>Mean</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-1.12</td>
<td>-2.25</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.78</td>
<td>0.81</td>
<td>0.84</td>
<td>1.12</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>UMN</td>
<td>Mean</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.53</td>
<td>-1.80</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.77</td>
<td>0.77</td>
<td>0.85</td>
<td>0.91</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>LI</td>
<td>Mean</td>
<td>0.00</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.77</td>
<td>0.78</td>
<td>0.89</td>
<td>0.97</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td>Cov(%)</td>
<td>96.4</td>
<td>94.1</td>
<td>95.2</td>
<td>94.3</td>
<td>94.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Y^S$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs</td>
<td>Mean</td>
<td>-0.01</td>
<td>0.26</td>
<td>-2.90</td>
<td>-5.08</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.79</td>
<td>0.82</td>
<td>0.83</td>
<td>1.06</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>IPW</td>
<td>Mean</td>
<td>0.01</td>
<td>-0.17</td>
<td>-1.25</td>
<td>-2.84</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.79</td>
<td>0.82</td>
<td>0.97</td>
<td>1.16</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>UMN</td>
<td>Mean</td>
<td>0.01</td>
<td>-0.15</td>
<td>-0.75</td>
<td>-2.38</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.79</td>
<td>0.82</td>
<td>0.89</td>
<td>1.12</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>LI</td>
<td>Mean</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.16</td>
<td>-0.98</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.79</td>
<td>0.82</td>
<td>0.93</td>
<td>1.20</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>Cov (%)</td>
<td>94.8</td>
<td>95.7</td>
<td>94.1</td>
<td>85.9</td>
<td>19.8</td>
</tr>
</tbody>
</table>

Table 1: Estimated mean responses and standard errors (SE) for Scenario 1 using observed data without correction for drop-out (Obs), 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>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>

Table 2: Estimated size and power of the diagnostic test, based on simulation results.
Figure 1: Boxplots of 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 \(Y^S\).

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

\[
Y_{a}^S = X_a \cdot B + S_a + \epsilon_a \quad Y_{a}^M = X_a \cdot B + M_a + \epsilon_a.
\]

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).

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 \(\pm 2\) empirical standard errors around the \(Y^M\) estimates. Bootstrap standard errors matched the empirical values closely. Standard errors derived from Proposition 1, which assumes negligible measurement error but avoids the need to bootstrap, 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).
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 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, 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
alternative 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 approaches are

A. Maximum likelihood estimation under a multivariate Normal model with unstructured covariance matrix (UMN). This approach assumes MAR and that missing data are ignorable.

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 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
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 in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>LI</th>
<th>OLS</th>
<th>A (UMN)</th>
<th>B (D+H)</th>
<th>C (IPW)</th>
<th>D (DYN)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(3.49)</td>
<td>(3.43)</td>
<td>(3.06)</td>
<td>(2.94)</td>
<td>(7.72)</td>
<td>(3.21)</td>
</tr>
<tr>
<td>P-S</td>
<td>13.04</td>
<td>6.01</td>
<td>11.01</td>
<td>13.66</td>
<td>12.37</td>
<td>12.42</td>
</tr>
<tr>
<td></td>
<td>(5.32)</td>
<td>(5.01)</td>
<td>(4.49)</td>
<td>(5.29)</td>
<td>(8.82)</td>
<td>(4.82)</td>
</tr>
<tr>
<td>E-S</td>
<td>-7.07</td>
<td>-1.43</td>
<td>-4.89</td>
<td>-5.97</td>
<td>-8.18</td>
<td>-5.40</td>
</tr>
<tr>
<td></td>
<td>(3.80)</td>
<td>(3.86)</td>
<td>(3.38)</td>
<td>(3.37)</td>
<td>(7.83)</td>
<td>(3.73)</td>
</tr>
</tbody>
</table>

means than the others, whereas method C adjusts almost as much as our dynamic linear 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. 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 \( t \). This is approach D above. Panel D of Figure 4 shows that the lack of fit of the original model, although easily detectable from the diagnostic plot, does not materially affect the fitted mean response profiles.

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, corresponding to a \( p \)-value of about 0.1. Together, the diagnostics suggest that departures from
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.
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 $Y$ 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. Indeed, one might argue that this is the only defensible approach to data analysis in the presence of drop-out.

In our view there is always an implicit time ordering. The analysis of longitudinal data, particularly when subject to missingness, should take this into account. 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 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.
What is therefore important is that all relevant information in $\mathcal{F}_t$ should be included in the model for the next expected increment. Figure 5 for example, 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 $\mathcal{Z}$ rather than $Y$ helps to preserve the interpretation of exogenous covariate effects. A modification of Proposition 3 could then be used to diagnose the suitability of these models for particular applications, though of course this could still only highlight potential problems in the observed data.

Using the language of stochastic processes, we have shown that 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’, in order to draw inference about the longitudinal features a population would have exhibited, assuming no-one had dropped out.

It would be perfectly feasible to consider more complicated random effects models for the increments of a longitudinal process, potentially gaining efficiency but requiring additional parametric assumptions. Such models could be thought of as analogous to frailty approaches to survival analysis (Vaupel et al., 1979), where unobservable random variables with a specified distributional form contribute to the local characteristics. Equally, alternative estimation procedures could be used: generalised estimating equations (Liang and Zeger, 1986) are one possibility. We have attempted to demonstrate that under certain specific assumptions and for certain specific objectives, our least squares procedure is a reasonable one. More complicated models for longitudinal data with drop-out would require similarly careful justification.
Appendix A: Fitting dynamic linear models using standard software

Least squares equations may be solved, and hence our linear increment models fitted, in virtually all software for statistical computing. We note, reflecting our own computing preferences, that this is particularly straightforward using the \texttt{lmList} command from the \texttt{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 \texttt{schizophrenia}, having columns \texttt{i} (a unique identifier), \texttt{time} (running from 1 to \(T_i\) for each \(i\)), \texttt{treat} (a factor indicating the treatment regime), and \texttt{PANSS}. This last column stores the change in PANSS associated with the given individual and time-point: that is, it contains \(\Delta Y_i(1), \ldots, \Delta Y_i(T_i)\) for every \(i\). Then

\[
> \text{fit} <- \text{lmList}(\text{PANSS} \sim \text{treat} \mid \text{time}, \text{data} = \text{schizophrenia}, \text{pool} = \text{F})
\]

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

\[
> \text{apply(coef(fit), 2, cumsum)}
\]

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

\[
> \text{SEs} <- \text{summary(fit)$coef[, "Std. Error", ]}
\]

may be extracted from the fitted model if, along the lines of Proposition 1, measurement error is thought to be negligible. These estimates (squared) may be summed

\[
> \text{apply(SEs}^2, 2, \text{cumsum})
\]

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

Appendix B: Proofs of Propositions 1 and 3

Proof of Proposition 1: We know that

\[
\hat{B}(t) - B^\mathcal{I}(t) = \sum_{s=1}^{t} \hat{\beta}(s) - 1_{\mathcal{I}}(s)\beta(s),
\]

and in the absence of measurement error, \(\hat{B} = B^\mathcal{I}\) is a martingale. Further, the predictable variation of a martingale is an unbiased estimator of its variance; the predictable variation of \(\hat{B} - B^\mathcal{I}\) evaluated at time \(t\) is

\[
\sum_{s=1}^{t} \text{V}(\hat{\beta}(s) - 1_{\mathcal{I}}(s)\beta(s)|\mathcal{F}_{s-}) = \sum_{s=1}^{t} \text{V}(X^{-}(s)\Delta M(s)|\mathcal{F}_{s-})
\]

\[
= \sum_{s=1}^{t} 1_{\mathcal{I}}(s)(X'(s)X(s))^{-1} \sigma_M^2(s).
\]
Just as in ordinary linear regression, $\tilde{\sigma}_M^2(t)$ (as defined in the statement of the proposition) is unbiased for $\sigma_\tau^2(t)$. 

**Proof of Proposition 3**: Let $Q = 1 \sigma (I - H)R$. Then
\[
\tilde{Z} = Q \cdot (M_a + \epsilon_a)
\]
since $M + \epsilon = R \cdot (M_a + \epsilon_a)$. For $t \geq 2$,
\[
E\{\tilde{Z}'(1)\tilde{Z}(t)\} = \text{tr} \left( E\{\tilde{Z}(1)\tilde{Z}'(t)\} \right)
\]
\[
= \text{tr} \left( E \sum_{s=1}^{t} Q(1) \{ \Delta M_a(1) + \Delta \epsilon'_a(1) \} \{ \Delta M'_a(s) + \Delta \epsilon'_a(s) \} Q(s) \right)
\]
\[
= \text{tr} \left( E \sum_{s=1}^{2} Q(1) E\{ \{ M_a(1) + \epsilon_a(1) \} \{ \Delta M'_a(s) + \Delta \epsilon'_a(s) \} | G_{s-} \} Q(s) \right)
\]
\[
= \text{tr} \left( E\{ Q(1) E\{ M_a(1) M'_a(1) + \epsilon_a(1) \epsilon'_a(1) | G_{1-} \} Q(1) - Q(1) \epsilon_a(1) \epsilon'_a(1) Q(2) \} \right)
\]
\[
= \text{tr} \left( E\{ Q(1) M_a(1) M'_a(1) Q(1) + Q(1) \epsilon_a(1) \epsilon'_a(1) \{ Q(1) - Q(2) \} \} \right)
\]
\[
= \text{scalar constant},
\]
as required. 

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, and 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


