A Martingale Residual Diagnostic for Longitudinal and Recurrent Event Data

Entisar Elgmati[1], Daniel Farewell[2] and Robin Henderson[1]

[1]Mathematics & Statistics, Newcastle University, UK
[2]Primary Care & Public Health, Cardiff University, UK

September 15, 2009

Abstract: One method of assessing the fit of an event history model is to plot the empirical standard deviation of standardised martingale residuals. We develop an alternative procedure which is valid also in the presence of measurement error and applicable to both longitudinal and recurrent event data. Since the covariance between martingale residuals at times t_0 and $t > t_0$ is independent of t, a plot of these covariances should, for fixed t_0, have no time trend. A test statistic is developed from the increments in the estimated covariances, and we investigate its properties under various types of model misspecification. Applications of the approach are presented using two Brazilian studies measuring daily prevalence and incidence of infant diarrhoea and a longitudinal study into treatment of schizophrenia.

Key words: covariance; dynamic covariate; event history; frailty; misspecification.

1 Introduction

A close connection exists between longitudinal and recurrent event data, since both are concerned with the evolution of random processes over time, and both are often subject to some form of censoring. Borgan et al (2007) describe an event-history approach to the analysis of longitudinal binary data arising from an investigation into incidence and prevalence of infant diarrhoea in Salvador, Brazil. Considering such data as discrete-time recurrent events brings at least three advantages. First, the powerful martingale machinery underpinning many event-history models facilitates inference, including covariate effect testing and standard error estimation. Second, it is very easy to incorporate dynamic effects by which previous individual-specific patterns can be built into intensity models (Aalen et al 2004, Fosen et al 2006). Third, estimation is extremely quick if, as in Borgan et al (2007), the Aalen additive intensity model is employed (Aalen 1980, 1989).

Being able to fit a model to complex data without time-consuming computation is important for good statistical practice. In rare cases, we may know exactly which model we are interested in, and computing time is not a major issue if only a single fit is needed. However, more often we need to fit many models, and here the advantages of quick computation

become apparent (Rue et al 2009). With many model fits comes the question of how best to compare models; Borgan et al (2007), following Aalen et al (2004), used informal inspection of plots of empirical standard deviations of standardised martingale residuals. Alternative tests, based on differenced martingale residuals, were described by Jones and Harrington (2001).

In this paper we propose a graphical procedure supplemented by formal tests. The paper extends an idea put forward by Diggle et al (2007) in the context of longitudinal data subject to dropout. Diggle et al used a martingale random effects approach for continuous longitudinal data and exploited the uncorrelated increments property for diagnostic purposes. A similar approach is developed here, the idea being to look at differences in covariances between martingale residuals at distinct time points. An advantage of this approach is that it is valid not just for recurrent event data but also for longitudinal data with subject-specific martingale random effects combined with additive measurement error.

In Section 2 we describe our diagnostic methods for general event-history models. Section 3 provides examples based on the Aalen intensity model and its extensions, and presents a simulation study. In Section 4 we turn to continuous responses. To emphasise some of the similarities, and differences, between longitudinal and event-history data, we examine how well our diagnostic can distinguish martingale from random slope models. Section 5 illustrates the proposed approaches using two sets of diarrhoea incidence and prevalence data and one set of longitudinal mental health data.

2 Model and martingale covariance diagnostic

Initially we will consider recurrent event data observed in continuous time. Adaptation to discrete time is straightforward (Borgan et al 2007). We assume independent censoring (Andersen et al 1993, p139) and also independence between subjects. The study period is from time zero to time $\tau < \infty$.

Let $N_i(t)$ count the number of events observed on subject $i = 1, 2, \ldots, n$ up to time t, and suppose there are (possibly time-varying) covariates $z_i(t) = (1, z_{i1}(t), z_{i2}(t), \ldots, z_{ip}(t))$. The intensity process of N_i is

$$\lambda_i(t) = \lambda_i(t|\mathscr{F}_{t-}) = Y_i(t)\alpha(t|z_i(t)),$$

where $\mathscr{F}_{t-}$ is the history of events, censoring, and covariates prior to time t. The process $Y_i(t)$ is an at-risk indicator, taking the value one when the individual may have an event observed, and zero otherwise. As usual the cumulative intensity is

$$\Lambda_i(t) = \int_0^t \lambda_i(u)du$$

and $N_i(t) = \Lambda_i(t) + M_i(t)$, where $M_i(t)$ is a martingale process with respect to $\mathscr{F}_t$.

Aalen et al (2004) describe a diagnostic plot based on standardised martingale residuals

$$M_i^*(t) = \frac{N_i(t) - \widehat{\Lambda}_i(t)}{\sqrt{\widehat{\mathrm{Var}}\{N_i(t)\}}} = \frac{\widehat{M}_i(t)}{\sqrt{\widehat{\mathrm{Var}}\{N_i(t)\}}}.$$

When the model is correctly specified, the empirical standard deviation (over subjects) of the $M_i^*(t)$ should be close to unity at all times t. A plot of such standard deviations at distinct event times should therefore be flat, and close to one. However, formal testing is awkward because of the strong dependence between the (estimated) standardised martingales at different event times. Instead, informal inspection of the plots is recommended and can provide valuable diagnostic information.

Diggle et al (2007), using ideas from event-history analysis, suggested a diagnostic for modelling continuous longitudinal data based on the idea that, since martingales have uncorrelated increments,

$$\mathrm{Cov}\{M_i(t_0), M_i(t)\} = \mathrm{Var}\{M_i(t_0)\} \tag{1}$$

when $0 \le t_0 < t$, for any fixed t_0. If the fitted model is correct, a plot of the left hand side of (1) evaluated at each time t should be flat. Departures from a horizontal line indicate non-suitability of the model being fitted.

Define $\overline{M}(t) = n^{-1}\{M_1(t) + \cdots + M_n(t)\}$, and let

$$C(t) = n^{-1}\sum_i \{M_i(t_0) - \overline{M}(t_0)\}M_i(t)$$

be the sample covariance of the true residuals of the observed data. We let $\widehat{C}(t)$ be the corresponding quantity based on estimated residuals $\widehat{M_i}$. Throughout, we assume t_0 to be held fixed. If the fitted model is correctly specified and estimators are $\sqrt{n}$-consistent, then the difference $\widehat{C}(u) - \widehat{C}(t) = O(n^{-1/2})$ for any $t_0 < t < u$.

Suppose now we are working either in discrete time or with a semiparametric model having non-zero intensity estimates only at distinct event times. In both cases, we can label the times of interest as $t_1 < \cdots < t_{k+m+1} = \tau$, setting $t_0 = t_k$ for some k. Define the m lag-1 differences in the covariance terms (after t_0) as $\Delta_j = C(t_{k+j+1}) - C(t_{k+j})$ for $j = 1, \ldots, m$, and the total difference as $\Delta_\bullet = \Delta_1 + \cdots + \Delta_m$, with $\widehat{\Delta}_j$ and $\widehat{\Delta}_\bullet$ denoting the corresponding sample-based estimates. Then, under a correctly specified model,

$$T = \widehat{\Delta}_\bullet \left(\sum_{j=1}^{m} \widehat{\Delta}_j^2 \right)^{-1/2} \tag{2}$$

is approximately standard Normal, provided the number of event times m (after t_0) is large. The proof, outlined in Appendix A, does not assume the $\widehat{\Delta}_j$ have equal variances, but does make use of asymptotic uncorrelation. There is no equivalent property for the standardised martingale residual procedure. We will turn to the question of how to test when m is not large in Section 4.

Our suggestion is thus to plot $\widehat{C}(t)$ against t for a given t_0, and to inspect for any obvious trend. This can be formalised by a test based on the test statistic T. We have not investigated optimal choice of t_0, but suggest it be chosen to be near — though not at — zero, so as to give a short time for processes to develop.

If the model is misspecified then $\widehat{C}(t)$ is consistent for

$$C^*(t) = E\{N(t_0)N(t) - \Lambda^*(t_0)N(t) - \Lambda^*(t)N(t_0) + \Lambda^*(t_0)\Lambda^*(t)\}$$

where $\Lambda^*(t)$ is the (so-called) least-false value to which the estimated intensity converges. This expression is used in our subsequent investigation of test power and other properties under model misspecification. Least-false values $\Lambda^*(t)$ and $C^*(t)$ for some representative models are given in Appendix B.

3 Properties

In order to assess in detail the properties of our proposed procedures, we must refer to specific models; we consider here a variety of semiparametric additive intensities. We use the usual method for fitting additive models (Borgan et al 2007), where increments in counting processes are regressed on the relevant covariates, whether dynamic or otherwise. We simulate in discrete time, with $t_j = j$ for $j = 1, \ldots, 100$, and $t_0 = t_{10}$. We have no censoring, and since events may recur, we have $Y_i(t) = 1$ for each i and all t.

We shall refer to the following models:

M1:	Two fixed covariates	$\alpha(t\|z) = \beta_0(t) + \beta_1(t)z_1 + \beta_2(t)z_2$
M2:	One fixed covariate	$\alpha(t\|z) = \beta_0(t) + \beta_1(t)z_1$
M3:	Dynamic D_∞	$\alpha(t\|z) = \beta_0(t) + \beta_1(t)z_1 + \beta_2(t)z_2 + \beta_3(t)D_\infty(t)$
M4:	Dynamic D_{30}	$\alpha(t\|z) = \beta_0(t) + \beta_1(t)z_1 + \beta_2(t)z_2 + \beta_3(t)D_{30}(t)$
M5:	Dynamic D_{20}	$\alpha(t\|z) = \beta_0(t) + \beta_1(t)z_1 + \beta_2(t)z_2 + \beta_3(t)D_{20}(t)$
M6:	Frailty	$\alpha(t\|z) = \min(Z\{\beta_0(t) + \beta_1(t)z_1 + \beta_2(t)z_2\}, 1)$

When simulating, we took z_1, z_2 to be independent standard Bernoulli variables. Since, in discrete-time models, intensities are also probabilities, we considered time-constant $\beta_j(t) = \beta_j$ that keep $\alpha(t|z)$ well below one.

The variables $D_u(t)$ are dynamic covariates defined as

$$D_u(t) = \begin{cases} \{N(t) - N(t-u)\}/u & t \geq u \\ N(t)/t & t < u. \end{cases} \tag{3}$$

We make the obvious definition that $D_\infty(t) = N(t)/t$ for all t.

The frailty variable Z is assumed to follow a gamma distribution, with mean one and variance ξ. Since the intensity depends on Z, this too may possibly exceed unity; hence the restriction to $[0, 1]$ in M6. Derivations in Appendix B take this restriction into account.

Table 1 shows the performance of the test statistic T when data are drawn from the frailty model M6, but models M1, M3, M4 and M5 are fitted to the data. The table includes results for various values of the frailty variance ξ; these include $\xi = 0$ under which M6 reduces to M1, so the test size can be checked. We used time-constant coefficients $\beta_0 = 0.1$, $\beta_1 = \beta_2 = 0.05$ for data generation but estimated them non-parametrically, as is customary for additive models. Sample sizes were $n = 250$ and $n = 500$, and 1,000 repetitions were used

to estimate size and power. The nominal test size was 5% and results show the empirical test size to be good and power to be very high when M1, M4 or M5 are fit, even for low values of frailty variance ξ. However, there is no power for concluding M3 is incorrect when the true model is M6. This is due to the strong similarity between frailty and dynamic models (Aalen et al, 2004).

Choice of the coefficients β in Table 1 gave an average of 15 events per subject over the follow-up period. Table 2 gives the rejection percentages when fitting the fixed-effects model M1 to data generated from the frailty model M6 with smaller β, and hence lower average numbers of events. Power decreases, as expected, but is still good.

Figure 1 illustrates the usefulness of $\widehat{C}(t)$ for model assessment. The figure is based on data generated from model M1 with two fixed covariates, fitting either the true model or the misspecified model M2, in which only one of the covariates is used. The diagnostic behaves as expected when the correct model is fitted, but shows a clear linear trend when the wrong model is employed. Indeed, a linear trend can be shown to be the least-false value for these circumstances (see Appendix B). The least-false value is shown as a solid line in the plot.

Least-false values for some other combinations are also given in Appendix B, and included in some of the panels of Figure 2. This figure shows the covariance diagnostic, and also the empirical standard deviation of standardised martingale residual processes (SMRP) when models M1, M3, M4 and M5 are fitted to frailty data generated under M6, with $\xi = 1$. The fixed-effects model M1 is clearly identified as being incorrect by both methods in these circumstances. However, the model with dynamic covariate D_∞ (M3) incorporating *all* previous history is not identified as being incorrect under either diagnostic. Again, this is because of the similarity between dynamic and additive frailty models. The covariance diagnostic identifies M4 and M5 as being incorrect much more clearly than does the SMRP. These models use only recent history in defining a dynamic covariate D_u (and, incidentally, are therefore correctly specified for the first u time units) whereas under the true frailty model M6 there is information in the full pattern of previous events.

When the frailty variance is smaller but not zero, the SMRP method tends to more clearly identify M4 and M5 as being incorrect. With small ξ, longer time periods are needed in order to identify between-subject heterogeneity and so there is less information in the more locally defined dynamic covariates.

4 Continuous responses model

The procedures described in Section 2 were motivated by a suggestion put forward by Diggle et al (2007) in the context of a martingale random effects model for continuous longitudinal data with dropout, whereas we have concentrated so far on recurrent event data. While the situations are similar, there are two notable differences. First, longitudinal models usually incorporate measurement error as well as subject-specific random effects. Second, typically a longitudinal study will have fewer potential observation times. Our asymptotic arguments (Appendix A) may therefore not apply. Hence in this section we examine performance for continuous longitudinal responses.

Let X_{ij} denote response at occasion j for subject i, with possibly time-varying covariates z_{ij}. We will take $t_0 = k$ and assume there are m further measurements so $j = 1, 2, \ldots, k + m$.

The Diggle et al linear increments model assumes a martingale random effect, in contrast to the ubiquitous Laird-Ware model with random slope and intercept. How easy is it to tell the two apart when there are rather few observations per subject?

We will consider a longitudinal model

$$X_{ij} = z_{ij}\beta_j + M_{ij} + \epsilon_{ij}, \tag{4}$$

where ϵ_{ij} denotes zero-mean measurement error and M_{ij} is a zero mean discrete time martingale. Independence between subjects is assumed together with within-subject independence of the ϵ_{ij}. We assume finite variances for both M_{ij} and ϵ_{ij}.

Since the number of distinct observation times $k+m$ may well be low for standard longitudinal studies, the assumed Normal distribution for T defined at (2) may not be valid. A valid alternative is to use a wild bootstrap method for non iid variables (Liu 1988, Mammen 1992), also known as conditional multiplier method (Martinussen and Scheike 2006, p43).

First note that we can write

$$\widehat{\Delta}_\bullet = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{M}_{ik}\{\widehat{M}_{i,j+k+1} - \widehat{M}_{i,j+k}\}$$

which estimates

$$\Delta_\bullet = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}M_{ik}\{M_{i,j+k+1} - M_{i,j+k}\}.$$

Now let $Q_1, Q_2, \ldots, Q_n$ be independent random variables with mean zero and unit variance. Since $E\{M_{i,j+k+1} - M_{i,j+k}\} = 0$ and the martingales M_i are mutually independent by assumption, Mammen (1992) shows that if $\sqrt{n}\Delta_\bullet$ converges in distribution to some variable D as n increases, then so too does

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}M_{ik}\{M_{i,j+k+1} - M_{i,j+k}\}Q_i.$$

Our assumptions of independence and finite moments are sufficient for the required convergence in distribution of $\Delta_\bullet$ to D. If we replace the true martingales by their estimators $\widehat{M}$ then there is no longer between-subject independence. However, the dependence and estimation error contributions are $O(n^{-1/2})$ and following arguments similar to Martinussen and Scheike (2006, 118-120), we can show that $\widehat{\Delta}_\bullet$ also converges in distribution to D.

We may also introduce predictable bounded weights $\{w_j\}$ to form a test statistic

$$\widehat{\Delta}^{(C)}_\bullet = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}w_j\widehat{M}_{ik}\{\widehat{M}_{i,j+k+1} - \widehat{M}_{i,j+k}\}.$$

For variance estimation we generate N replications of n standard Normal random variables $Q_1, Q_2, \ldots, Q_n$. For each replication the conditional multiplier version of $\widehat{\Delta}^{(C)}_\bullet$ is calculated

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}w_j\widehat{M}_{ik}\{\widehat{M}_{i,j+k+1} - \widehat{M}_{i,j+k}\}Q_i$$

and the empirical variance of these consistently estimates the variance of $\widehat{\Delta}_{\bullet}^{(C)}$. In the application and simulations to come we shall use two choices of weights w_j. The first gives equal weight to all time points and so tests for overall change in covariances:

$$\widehat{\Delta}_{\bullet 1}^{(C)} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \widehat{M}_{ik} \{ \widehat{M}_{i,j+k+1} - \widehat{M}_{i,j+k} \}. \tag{5}$$

The second gives linearly increasing weights and so tests for trend:

$$\widehat{\Delta}_{\bullet 2}^{(C)} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} j \widehat{M}_{ik} \{ \widehat{M}_{i,j+k+1} - \widehat{M}_{i,j+k} \}. \tag{6}$$

This may be appropriate when we have under consideration an alternative Laird-Ware random effects model

$$X_{ij} = z_{ij}\beta_j + U_{i0} + U_{i1} t_j + \epsilon_{ij}, \tag{7}$$

where U_{i0} and U_{i1} are possibly correlated zero mean Normal random intercept and slope respectively. If (7) is true but (4) is fitted then the least false values $C^*(t)$ are linear in t (to first order) and hence the slope of the empirical values is suggested as test statistic. An exception is when $t_0 = 0$ ($k = 1$) and U_{i0} and U_{i1} are independent, in which case $C^*(t)$ is constant in t (again, to first order) so diagnostics will have little or no power.

We use a simple simulation to investigate performance of the three suggested tests: statistic T defined at (2), and $\widehat{\Delta}_{\bullet 1}^{(C)}$ and$\widehat{\Delta}_{\bullet 2}^{(C)}$ defined at (5)and (6) respectively. Following Diggle et al (2007) we assume there are no covariates and measurements are scheduled at times $(t_j : j = 1, \ldots, 6) = (0, 1, 2, 4, 6, 8)$. We take $\mathrm{Var}(U_{i0}) = 200, \mathrm{Var}(U_{i1}) = 200, \mathrm{Cov}(U_{i0}, U_{i1}) = 0$ and $\mathrm{Var}(\epsilon_{ij}) = 100$. Increments in the martingale of (4) are independent Normal variables chosen so that $\mathrm{Var}(M_{ij}) = \mathrm{Var}(U_{i0} + U_{i1} t_j)$. There is no censoring.

Table 3 gives results. Due to poor variance estimation, the test statistic T performs very poorly, with no rejections under any of the combinations given. Tests $\widehat{\Delta}_{\bullet 1}^{(C)}$ and$\widehat{\Delta}_{\bullet 2}^{(C)}$ perform similarly with correct test size when the martingale model (4) is true and high power when (7) is true and $t_0 = t_2 = 1$ is selected as base time. As expected, the tests have no power at $t_0 = t_1 = 0$ since the expected value of the covariance estimate is constant, even under (7).

5 Applications

5.1 Recurrent event data: incidence and prevalence of infant diarrhoea

We illustrate the use of our diagnostics with two data sets recording daily diarrhoea events in young children.

The first data set, from the *Blue Bay* study, was also considered by Borgan et al (2007). Daily data are available from 926 children monitored between October 2000 and January 2002 in Salvador, Brazil. For each child, we have longitudinal binary data, or equivalently discrete-time recurrent event data, recording whether or not they experienced diarrhoea each day. All children were under 3 years of age at recruitment. Prevalence was 5% at the

beginning of the study, falling to 1% by the end. Incidence was 2% to begin with and 0.5% at the close. A variety of covariate information is available, relating to living conditions, local environment, family circumstances and so on. We shall not discuss these here, nor the complex patterns of missing data: see Borgan et al (2007) for further information. Instead, we concentrate on diagnostics.

We consider both incidence and prevalence and compare models with and without the inclusion of dynamic covariates. A prevalence event occurs if a child has diarrhoea during day j, an incidence event occurs if a child begins a new episode of diarrhoea during day j. An episode is a sequence of days with diarrhoea until there have been three diarrhoea-free days. For both types of events we fitted discrete-time additive intensity models

$$\mathrm{P}(\text{event on day } j|\mathscr{F}_{t_j-}) = Y_j\{\beta_{0j} + \beta_{1j}\, z_{i1j} + \cdots + \beta_{pj}\, z_{ipj}\} \tag{8}$$

where $\mathscr{F}_{t^-}$ denotes history up to but not including time t, Y_j is a predictable at-risk indicator and $(z_{i1j}, \ldots, z_{ipj})$ are possibly time-varying covariates. These are made up of a mixture of baseline covariates, age on day j, and dynamic covariates summarising events over previous days. Baseline covariates for the Blue Bay data were all indicator variables, recording gender and whether: there were three or more people per bedroom; the street was of low quality; the water storage was contaminated; the water source was contaminated; there was standing water; there was open sewerage; the accommodation was rain-affected; the mother was less than 25 years of age; the family was of low socio-economic status; or there were other children aged less than 5 years in the household. Dynamic covariates were the previous rate (events divided by days at risk) of diarrhoea, fever or vomit, and for the prevalence model whether or not there was diarrhoea on the immediately previous four days.

We will not report results in full here. Instead we give the diagnostic plots in Figure 3. Recall that for a correctly specified model we expect the SMRP plot to be close to one at all times, and that the covariance process $\widehat{C}(t)$ should be constant. The plot shows both SMRP and $\widehat{C}(t)$ for the prevalence and incidence analyses. Each panel has three lines, corresponding to fitting a model without dynamic covariates, with only the most significant dynamic covariate also included, and with all dynamic covariates included. For the prevalence analysis the apparently most important dynamic covariate was the rate of diarrhoea events, defined as number of previous days with diarrhoea scaled by days at risk. For the incidence analysis the most important dynamic covariate was the number of previous episodes, again scaled by risk days. It is clear that inclusion of the dynamic covariates substantially improves the model fits for both prevalence and incidence, and that there is additional gain by including all dynamics, not just one. We can supplement this conclusion by the test T defined at (2). P-values are given in Table 4: there is strong evidence against the models that ignore previous events and assume baseline/fixed covariates only (including age), whereas there is no such evidence of misspecification once the dynamic covariates are included.

The second data set has a similar structure, but relates to a double-blind placebo controlled trial carried out in the town of *Serrinha* in the state of Bahia in North East Brazil (Barreto et al, 1994). We have daily data for 1,192 children aged between 6 and 48 months at recruitment and followed for one year. Prevalence and incidence overall were around 5% and 2% respectively. Fixed covariates included gender, floor type, source of water, presence of toilet, treated water, open sewerage, other children in the household and level of mothers education. In addition, for this study we have an indicator for whether or not the child received vitamin A supplements. Dynamic covariates included all those used in the Blue

Bay analysis as well as the proportion of previous days during which the child had other diseases.

Again we will omit results of model fitting, other than diagnostics. These are given in Figure 4 and Table 4. Conclusions are similar to those for Blue Bay: without the dynamic covariates there is clear model misspecification, whereas there is no real evidence against the models that include dynamic covariates. In this case, the most significant dynamic covariates (day and episode rate) may perhaps not need to be supplemented by the vomit, fever and other disease rate measures.

5.2 Continuous response data: a longitudinal study of mental health

Henderson et al (2000) describe a clinical trial in which 518 subjects in three treatment groups were scheduled to have measurements of mental health measured longitudinally over 8 weeks. Patients were all diagnosed with schizophrenia and the treatment groups were placebo, a standard drug and an experimental drug. The measurement of mental health was the positive and negative symptom scale (PANSS), which was scheduled to be obtained at weeks 0,1, 2, 4, 6 and 8. However, patients dropped out throughout the trial and only 269 completed it.

Diggle et al (2007) reanalysed these data using a dynamic linear increments model, which can be written in the form

$$X_{it} = (Z_i \cdot B)_t + M_{it} + \epsilon_{it},$$

where $(Z_i \cdot B)$ is a transform and M_{it} is a martingale, both in discrete time. We will concentrate here on diagnostics for the martingale assumption: see Diggle et al (2007) for further information on estimation and dealing with dropout.

The estimated covariances between residuals at week 0 ($= t_0$) and those at weeks 1–8 (with bootstrap standard errors) are:

	Week t				
	1	2	4	6	8
$\widehat{C}(t)$	247.91	239.38	234.42	228.16	217.61
ESE	16.44	18.12	18.79	19.51	20.10

Standardised test statistics $\widehat{\Delta}^{(C)}_{\bullet 1}$ are given in Table 5 for all combinations of baseline t_0 and final week t. Trend tests $\widehat{\Delta}^{(C)}_{\bullet 2}$ are very similar and hence omitted. Although there is some evidence that the martingale assumption is becoming questionable by the final week, none of the tests are individually significantly different from zero and our conclusion is that this form of subject-specific martingale error process is plausible for these data. Note that we have not made any assumptions about homogeneity of variance either over time or between subjects.

6 Discussion

One way of assessing the fit of an event-history model is to plot the empirical standard deviation of the standardised martingale residual processes

$$M_i^*(t) = \frac{\widehat{M}_i(t)}{\sqrt{\widehat{\mathrm{Var}}\{N_i(t)\}}}.$$

This gives the SMRP plots we have illustrated in simulations and two applications. We propose such plots be supplemented by, rather than replaced with, plots of estimated martingale covariances

$$\widehat{C}(t) = n^{-1} \sum_i \widehat{M}_i(t_0)\widehat{M}_i(t).$$

The SMRP procedure has the advantage that we know the estimates should be close to one under the model, whereas all we know for the covariance diagnostic is that there should be no trends with time. On the other hand, as pointed out by a referee, any lack of fit for SMRP could be explained by either the model being a correct rate model but the standard error not being estimated correctly, or the model being a correct rate model. The covariance diagnostic procedure does not incorporate a model-based standard error estimate and hence is robust to this aspect of misspecification. Further, the SMRP procedure can be problematic when, as can happen under an additive model, the intensity estimates are occasionally negative or, for discrete-time data, larger than one. In these cases the estimated variance can occasionally be negative, so some smoothing is needed in order to use $M_i^*(t)$. The covariance diagnostic also has the advantage of asymptotically independent increments and hence the availability of the test statistics T, $\widehat{\Delta}_{\bullet 1}^{(C)}$ and $\widehat{\Delta}_{\bullet 2}^{(C)}$ defined at (2),(5) and (6). Moreoever, because it need not be scaled by a variance estimate, computing expected patterns under misspecification is more tractable for the covariance diagnostic (Appendix B).

A referee raised the question of what type of misspecification can be infereed from any particular non-null pattern in the plots, given a significant test statistic. Figures 1–2 illustrate some possibilities but unfortunately the type of misspecification away from the null is not identifiable using our procedures under the semiparametric approach with covariates allowed to have unspecified time-varying effects. An omitted additive covariate z_j could easily lead to the same type of summary plot as an omitted multiplicative frailty term for instance, if the true regression effect $\beta_j(t)$ has the right form. This is not to say that further investigation is not worthwhile: we view our procedures as omnibus goodness of fit checks, to be followed by further detailed investigation if the null model is rejected. Thus we might go on for instance to use the cumulative sums of martingale residuals to check functional form of covariates (Lin et al 1993).

We have not explored choice of reference time t_0 for use in $\widehat{C}(t_0)$. If too small a t_0 is chosen, then $\widehat{M}(t_0)$ may be unstable, or result in a constant least-false value (see Section 4). Of course, choosing too large a t_0 results in less power to detect misspecification. We could also plot $\widehat{C}(t)$ for $t < t_0$, to attempt to provide some verification of model adequacy over these earlier times. The option of trying several t_0 as in Table 5 might also be fruitful, but has not yet been thoroughly investigated. Similarly, we have not investigated here the performance of $\widehat{C}(t)$ for single event survival data. This, too, is a topic for ongoing work in the cross-pollination of methods between longitudinal and event-history data.

Acknowledgement

Our thanks go to Rosemeire Fiaccone and Mauricio Barreto for providing the diarrhoea data studied in Section 5. We thank also an associate editor and two anonymous referees.

Appendix A: Justification of test statistic

We give an outline proof of the asymptotic Normality of test statistic (2). Assuming that the true martingales are locally square integrable a fully rigorous proof is available.

Define the lag-1 differences in the covariance terms as

$$\Delta_j = C(t_{k+j+1}) - C(t_{k+j}) \qquad \text{and} \qquad \widehat{\Delta}_j = \widehat{C}(t_{k+j+1}) - \widehat{C}(t_{k+j}).$$

We will consider the true martingale residuals first. We need to show first that the $\widehat{\Delta}_j$ are asymptotically uncorrelated. It will suffice to show that $\mathrm{Cov}(\Delta_j, \Delta_{j+1}) = 0$ for $j = 1, \ldots, m-1$. Note that

$$\Delta_j = n^{-1} \sum_i \{M_i(t_0) - \overline{M}(t_0)\}\{M_i(t_{k+j+1}) - M_i(t_{k+j})\}.$$

Since martingale increments are uncorrelated — specifically, since $\{M_i(t_{k+j+1}) - M_i(t_{k+j})\}$ is not correlated with $\{M_i(t_{k+j+2}) - M_i(t_{k+j+1})\}$ — we have $\mathrm{Cov}(\Delta_j, \Delta_{j+1}) = 0$ for $j = 1, \ldots, m-1$.

Since we assume our estimator is $\sqrt{n}$-consistent (i.e. $\widehat{M}(t) = M(t) + O(n^{-1/2})$), the same result holds to first order when we replace $M(t)$ with its estimate $\widehat{M}(t)$: we have $\mathrm{Cov}(\widehat{\Delta}_j, \widehat{\Delta}_{j+1}) = O(n^{-1/2})$.

Noting that $\widehat{\Delta}_j$ has expectation zero for a correctly specified model and that the covariances decrease to zero with n, it follows from the Central Limit Theorem that $m^{-1}\widehat{\Delta}_\bullet = m^{-1}(\widehat{\Delta}_1 + \cdots + \widehat{\Delta}_m)$ is asymptotically zero mean Gaussian with variance consistently estimated by

$$\sum_{j=1}^{m} \widehat{\Delta}_j^2 / m^2$$

as the number of event times m increases.

Appendix B: Properties under misspecification

To illustrate the calculation of the expected shape of the SMRP and covariance diagnostic plots under model misspecification, we consider the case of an ignored covariate. Assume that the true model is given by

$$\lambda_i(t) = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2}$$

where z_{i1} and z_{i2} are two time constant binary independent random variables. For the analysing model we will use z_1 as the only covariate, and see what will happen to the

variance of the standardised residual processes. Consider to begin with

$$\begin{aligned}\lambda_i^* = E\{\lambda_i(t)|z_{i1}\} &= \beta_0 + \beta_1 z_{i1} + \beta_2 E(z_{i2}|z_{i1}) \\ &= \beta_0 + \beta_1 z_{i1} + \beta_2 \times \tfrac{1}{2},\end{aligned}$$

by the independence of z_1 and z_2. As sample size increases, we shall have $\widehat{\lambda}_i(t) \to \lambda_i^*$ (and $\widehat{\Lambda}_i(t) \to \lambda_i^* t$).

Let S_n be the sample variance of the estimated standardised residual processes $M_i^*(t)$. Since

$$M_i^*(t) = \frac{N_i(t) - \widehat{\Lambda}_i(t)}{\sqrt{\sum_{t_j \le t} \widehat{\lambda}_i(t_j)\{1 - \widehat{\lambda}_i(t_j)\}}} \to \frac{N_i(t) - \lambda_i^* t}{\sqrt{\sum_{t_j \le t} \lambda_i^*\{1 - \lambda_i^*\}}},$$

the sample variance S_n will be (approximately, for large n)

$$S_n(t) = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{N_i(t) - \lambda_i^* t}{\sqrt{\sum_{t_j \le t} \lambda_i^*\{1 - \lambda_i^*\}}} \right]^2$$

For notational ease we now drop the subscript i. By the law of large numbers, $S_n(t)$ will approach the mean of its summand as n increases. Now the summand of this expression is a function of three random variables: z_1, z_2 and $N(t)$. Therefore $E\{S_n(t)\} = E[E\{S_n(t) \mid z_1, z_2\}]$, the inner expectation of which equals

$$E\left[\frac{\{N(t) - \lambda^* t\}^2}{\sum_{t_j \le t} \lambda^*\{1 - \lambda^*\}} \mid z_1, z_2 \right].$$

The denominator is a constant (given z_1 and z_2), while the numerator can be expanded as $[\{N(t) - \Lambda(t)\} - \{\lambda^* t - \Lambda(t)\}]^2$. Now $E(\{N(t) - \Lambda(t)\}^2 \mid z_1, z_2) = \mathrm{Var}(N(t) \mid z_1, z_2)$, $E(N(t) - \Lambda(t) \mid z_1, z_2) = 0$, and $\lambda^* t - \Lambda(t) = \beta_2 t(\frac{1}{2} - z_2)$. Recalling that we will be simulating in discrete time,

$$\begin{aligned}\mathrm{Var}\{N(t) \mid z_1, z_2\} &= \sum_{t_j \le t} \lambda(t_j)\{1 - \lambda(t_j)\} \\ &= (\beta_0 + \beta_1 z_1 + \beta_2 z_2)\{1 - (\beta_0 + \beta_1 z_1 + \beta_2 z_2)\} t.\end{aligned}$$

Thus

$$E\{S_n(t) \mid z_1, z_2\} = \frac{(\beta_0 + \beta_1 z_1 + \beta_2 z_2)\{1 - (\beta_0 + \beta_1 z_1 + \beta_2 z_2)\} t + \beta_2^2 t^2 (\frac{1}{2} - z_2)^2}{(\beta_0 + \beta_1 z_1 + \frac{1}{2}\beta_2)\{1 - (\beta_0 + \beta_1 z_1 + \frac{1}{2}\beta_2)\} t}.$$

Now taking expectation with respect to z_2 and z_1, which we recall are binary, we can show that

$$E(S_n(t)) = 1 + \tfrac{1}{8}\beta_2^2 t \left\{ \frac{1}{(\beta_0 + \frac{1}{2}\beta_2) - (\beta_0 + \frac{1}{2}\beta_2)^2} + \frac{1}{(\beta_0 + \beta_1 + \frac{1}{2}\beta_2) - (\beta_0 + \beta_1 + \frac{1}{2}\beta_2)^2} \right\}.$$

Hence we expect the SMRP plot to be have a non-zero linear trend with t if a covariate is omitted, and the true regression coefficients are time-constant.

A similar argument for the covariance diagnostic leads to least-false value

$$\widehat{C}(t) = \tfrac{1}{4}\left[t t_0 \beta_2^2 + (4\beta_0 + 2\beta_1 + 2\beta_2) t_0 - \{(\beta_0 + \beta_1 + \beta_2)^2 + (\beta_0 + \beta_2)^2 + (\beta_0 + \beta_1)^2 - \beta_0^2\} t_0^2 \right]$$

which is also linear in t, as seen in Figure 1.

Least false values when either fixed effects or dynamic models are fitted to data generated with frailty can also be calculated (Elgmati, 2009).

References

Aalen, O O (1980). A model for non-parametric regression analysis of counting processes. In Klonecki,W, Kozek, A and Rosinski, T, editors, Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory, Springer-Verlag New York, 1-25.

Aalen, O O (1989). A linear regression model for the analysis of life times. Statistics in Medicine, **8**, 907-925.

Aalen, O O (1993). Further results on the non-parametric linear regression model in survival analysis. Statistics in Medicine, **12**, 1569-1588.

Aalen, O O, Fosen, J, Wedon-Fekjær, H, Borgan, Ø, and Husebye, E (2004). Dynamic analysis of multivariate failure time data. Biometrics, **60**, 764-773.

Albert, P S (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. Biometrics **56**, 602-608.

Andersen, P K, Borgan, Ø, Gill, R d and Keiding, N (1993). Statistical Models Based on Counting Processes. Springer-Verlag, New York.

Barreto, M L, Santos, L M, Assis, A M, Araũjo, M P, Farenzena, G G, Santos, P A, Fiaccone, R L (1994). Effect of vitamin A supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in Brazil. Lancet, **344**, 228-231.

Borgan Ø., Fiaccone R L, Henderson R and Barreto M L (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. Scandinavian Journal of Statistics, **34**, 53-69.

Diggle, P, Farewell, D, Henderson, R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal (with discussion) (2007) . Journal of the Royal Statistical Society: Series C (Applied Statistics) **56**, 499-550.

Elgmati, E. (2009). Additive intensity models for discrete time recurrent event data. Unpublished PhD thesis, Newcastle University, UK.

Fosen, J, Borgan, Ø, Wedon-Fekjær, H and Aalen, O O (2006). Dynamic analysis of recurrent event data using the additive hazard model. Biometrical Journal, **48**, 381-398.

Henderson, R, Diggle, P J and Dobson, A (2000). Joint modelling of longitudinal measurements and event time data. Biostatistics, 1, 465480.

Hjort, N L (1992). On inference for parametric survival data models. International Statistical Reviews, 60, 355-387.

Lin, D Y, Wei, L J and Ying, Z (1993). Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika, 80, 557-572.

Liu, R Y (1988). Bootstrap procedures under some non iid models. Annals of Statistics, 16, 1696-1708.

Mammen, E (1992). Bootstrap, wild bootstrap, and asymptotic normality. Probability Theory and Related Fields, 93, 439-455.

Marinussen, T and Scheike, T H (2006). Dynamic Regression Models for Survival Data. Springer, New York.

Rue, H, Martino, S and Chopin, H (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B, to appear.

Jones, C L and Harrington, D P (2001). Omnibus tests of the martingale assumption in the analysis of recurrent event failure time data. Lifetime Data Analysis, 7, 157-172.

Strina, A, Cairncross, S, Prado, M S, Teles, C A and Barreto, M L (2005). Childhood diarrhoea symptoms, management and duration: observations from a longitudinal community study. Transactions of the Royal Society of Tropical Medicine and Hygiene, 99, 407-416.

Table 1: Estimated test power when fitting M_1, M_3, M_4 or M_5 to data generated under frailty model M6 with frailty variance ξ. At $\xi = 0$ model M6 is equivalent to model M1. Results are from 1000 simulated samples with $\beta_0 = 0.1, \beta_1 = \beta_2 = 0.05$.

n	$n = 250$					$n = 500$				
Model \ ξ	0.0	0.05	0.1	0.5	1	0.0	0.05	0.1	0.5	1
M_1	5%	78%	98%	100%	100%	5%	95%	100%	100%	100%
M_3	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%
M_4	17%	67%	86%	100%	100%	32%	90%	99%	100%	100%
M_5	24%	83%	97%	100%	100%	43%	98%	100%	100%	100%

Table 2: Effect of number of events on power of diagnostic test when fitting fixed effects model M1 to data generated from frailty model M6. Results are from 1000 samples of size 500.

ξ	$\beta_0 = 0.1$ $\beta_1 = \beta_2 = 0.05$	$\beta_0 = 0.05$ $\beta_1 = \beta_2 = 0.025$	$\beta_0 = 0.05$ $\beta_1 = \beta_2 = 0.0125$	$\beta_0 = 0.025$ $\beta_1 = \beta_2 = 0.0125$	$\beta_0 = 0.025$ $\beta_1 = \beta_2 = 0.00833$
1	100%	100%	100%	100%	100%
0.5	100%	100%	100%	100%	100%
0.25	100%	100%	100%	96%	92%
0.125	100%	97%	92%	60%	51%
0.09	100%	88%	75%	38%	33%
0.07	100%	72%	57%	27%	21%
0.05	95%	49%	35%	17%	13%
0.03	70%	25%	18%	10%	9%
0.009	14%	7%	6%	6%	5%
0.007	12%	7%	6%	6%	5%
0.005	8%	5%	6%	5%	5%
0.0	5%	5%	5%	5%	5%

Table 3: Test sizes and power for continuous response simulation scenario of Section 4. $T =$ test of (2) in Section 2; $\widehat{\Delta}^{(C)}_{\bullet 1}, \widehat{\Delta}^{(C)}_{\bullet 2}$ = conditional multiplier tests (5)and (6) of Section 4.

		$t_0 = 0$			$t_0 = 1$		
True model	Test	$n = 100$	$n = 250$	$n = 500$	$n = 100$	$n = 250$	$n = 500$
Martingale	T	0.000	0.000	0.000	0.000	0.000	0.000
	$\widehat{\Delta}^{(C)}_{\bullet 1}$	0.041	0.056	0.038	0.051	0.045	0.051
	$\widehat{\Delta}^{(C)}_{\bullet 2}$	0.043	0.057	0.046	0.050	0.048	0.051
Random slope	T	0.000	0.000	0.000	0.000	0.000	0.000
	$\widehat{\Delta}^{(C)}_{\bullet 1}$	0.042	0.055	0.048	0.446	0.818	0.988
	$\widehat{\Delta}^{(C)}_{\bullet 2}$	0.040	0.049	0.053	0.411	0.790	0.986

Table 4: P-values for covariance test (3) for Blue Bay and Serrinha data.

		Prevalence	Incidence
Data	Model	p	p
Blue Bay	Fixed covariates only	0.0000	0.0000
	Dynamic also included	0.5507	0.7427
Serrinha	Fixed covariates only	0.0000	0.0006
	Dynamic also included	0.6488	0.9426

Table 5: Test statistics $\widehat{\Delta}_{\bullet 1}^{(C)}$ for schizophrenia data. Values given are standardised by the conditional multipler standard deviation estimate obtained from 100 replications

t_0	t			
	2	4	6	8
0	-0.67	-0.91	-1.13	-1.63
1		-0.91	-1.47	-1.93
2			-0.80	-1.66
4				-1.51

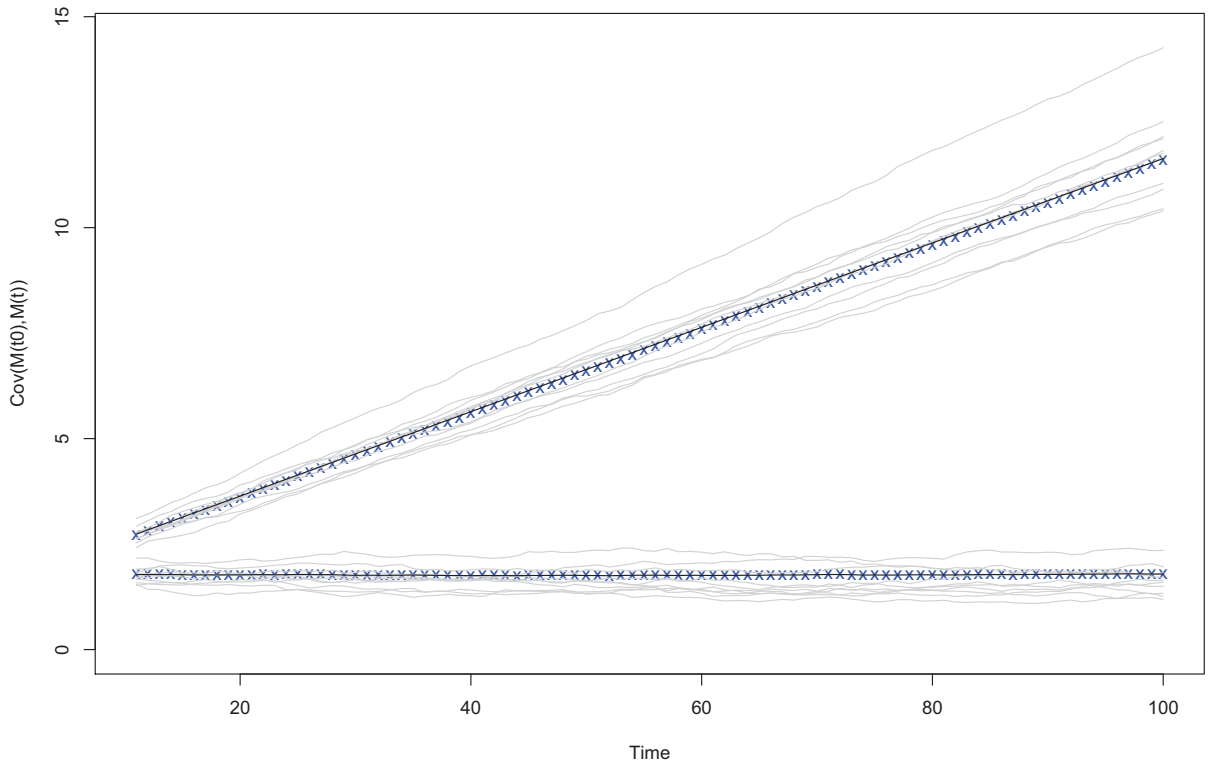

Figure 1: Covariance diagnostic for data generated under M1 and fitted using M1 (lower lines) or M2 (upper lines). Means of 100 samples of sample size 500 are shown as points. The grey lines show results from 10 randomly chosen single simulated data sets. In both parts of the plot the theoretical least-false value is included as a solid line.

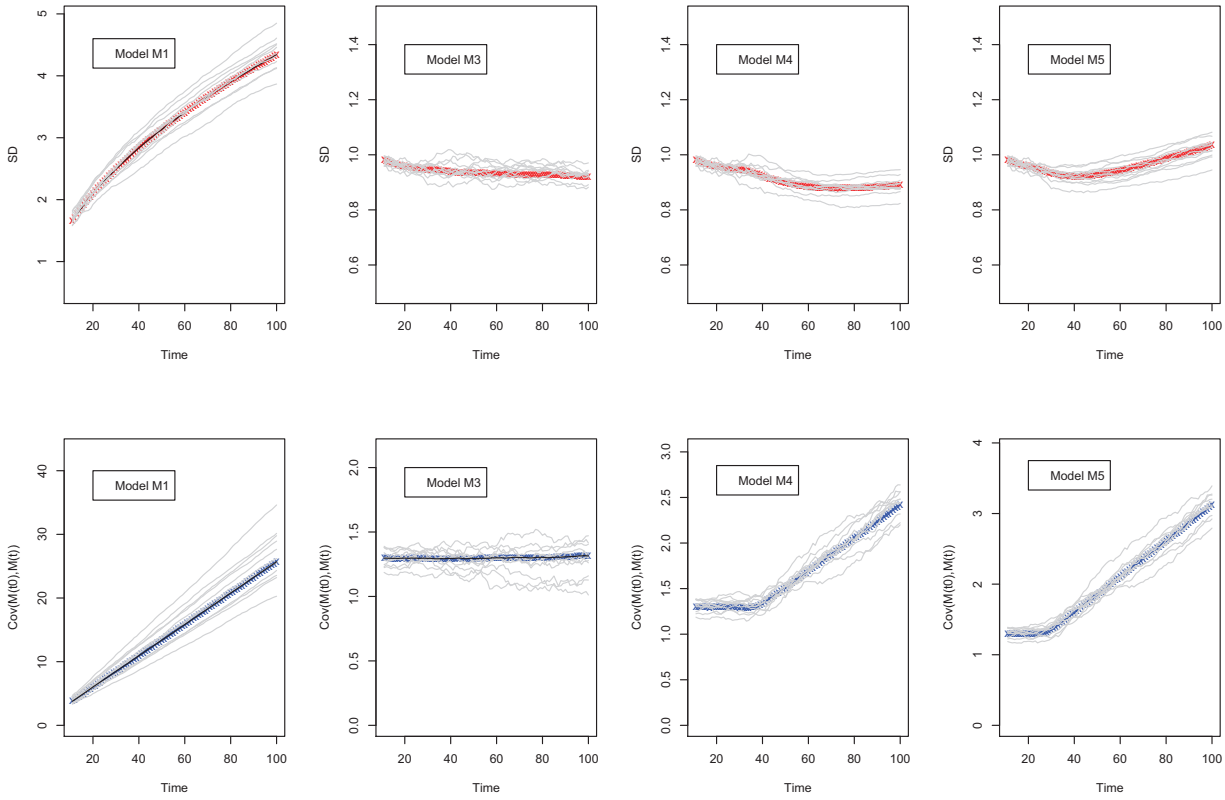

Figure 2: Covariance (lower panels) and SMRP (upper panel) diagnostic plots for data generated under frailty model M6 with $\xi = 1$ and fitted using $M1$, $M3$, $M4$ and $M5$. The grey lines show results from 10 randomly chosen single simulated data sets. Plots M1 for SMRP and M1 and M2 for the covariance include as solid lines the theoretical least-false values.

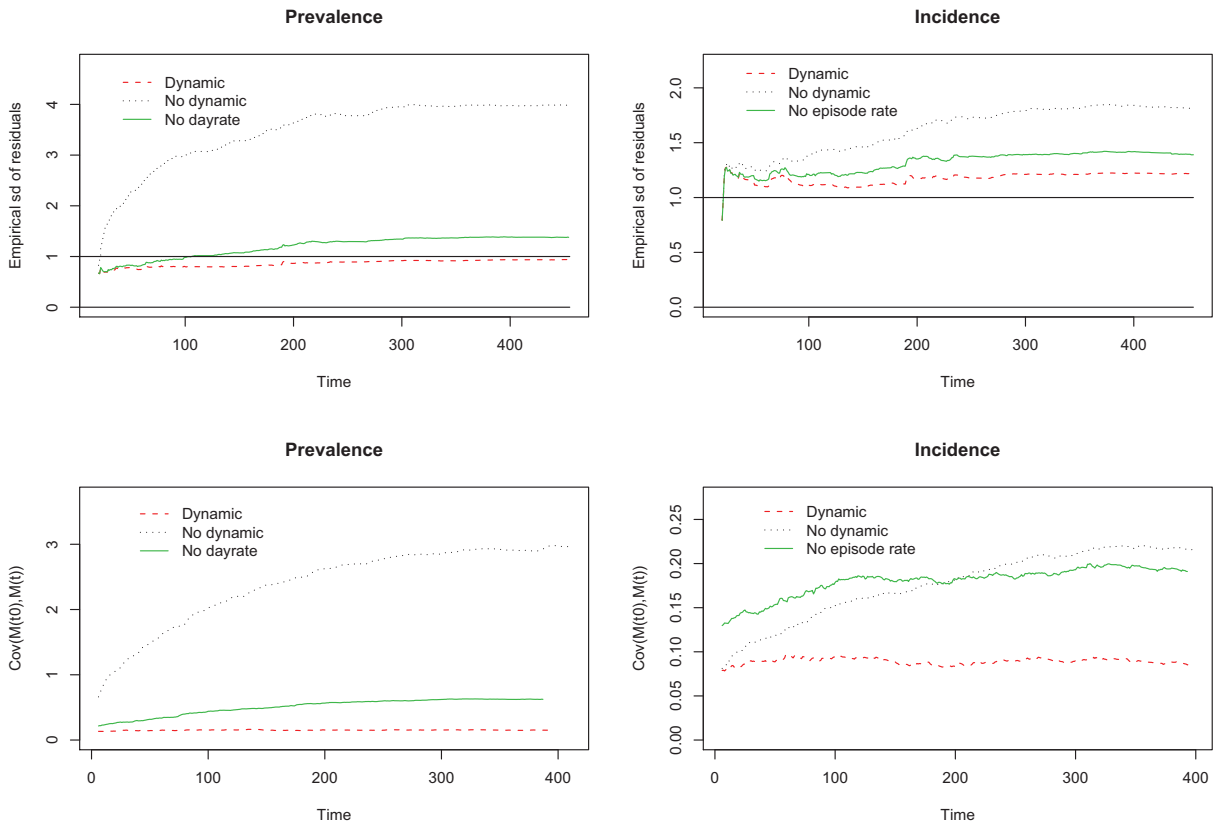

Figure 3: SMRP (upper panels) and covariance (lower panels) diagnostic plots for additive model fits to Blue Bay data, with and without inclusion of dynamic covariates.

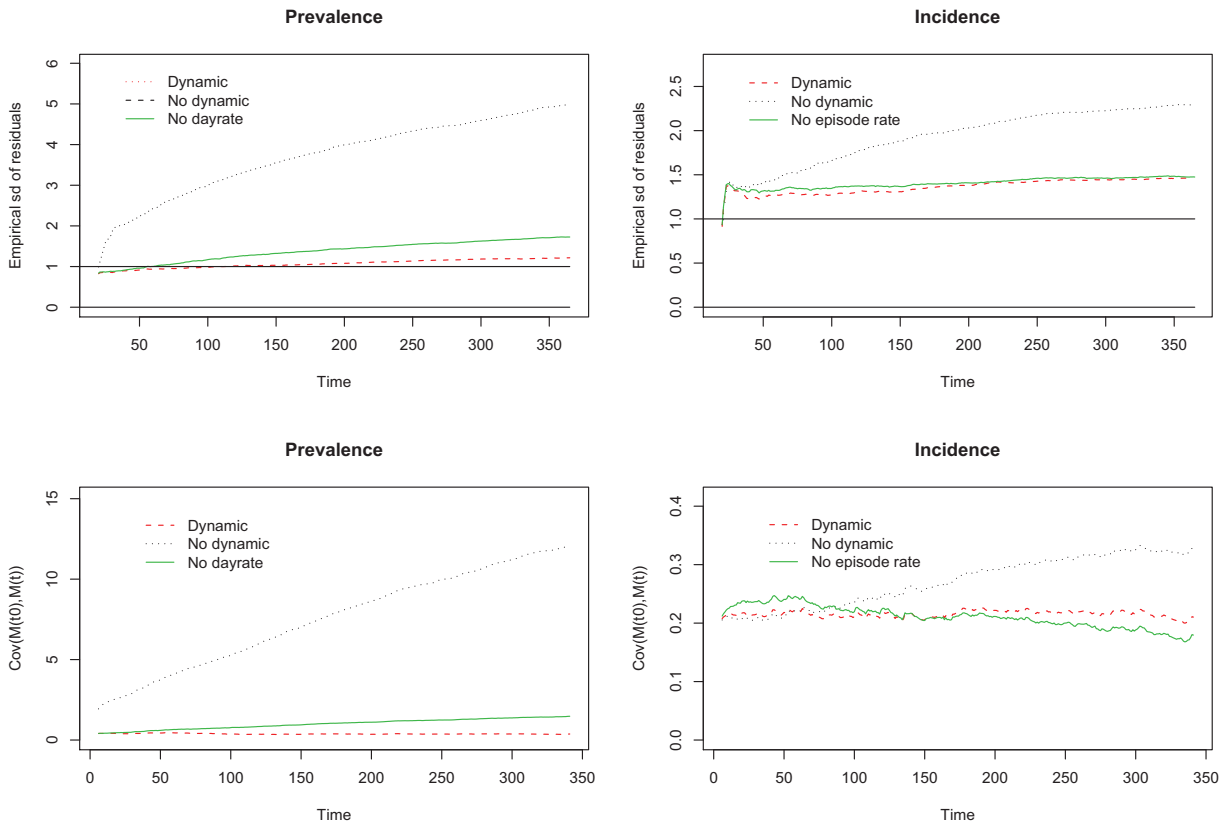

Figure 4: SMRP (upper panels) and covariance (lower panels) diagnostic plots for additive model fits to Serrinha data, with and without inclusion of dynamic covariates.