Lifetime Data Anal (2015) 21:542–560
DOI 10.1007/s10985-015-9321-4

Penalised logistic regression and dynamic prediction for discrete-time recurrent event data

Entisar Elgmati · Rosemeire L. Fiaccone ·
R. Henderson · John N. S. Matthews

Received: 18 August 2014 / Accepted: 7 January 2015 / Published online: 28 January 2015

Abstract We consider methods for the analysis of discrete-time recurrent event data, when interest is mainly in prediction. The Aalen additive model provides an extremely simple and effective method for the determination of covariate effects for this type of data, especially in the presence of time-varying effects and time varying covariates, including dynamic summaries of prior event history. The method is weakened for predictive purposes by the presence of negative estimates. The obvious alternative of a standard logistic regression analysis at each time point can have problems of stability when event frequency is low and maximum likelihood estimation is used. The Firth penalised likelihood approach is stable but in removing bias in regression coefficients it introduces bias into predicted event probabilities. We propose an alterative modified penalised likelihood, intermediate between Firth and no penalty, as a pragmatic compromise between stability and bias. Illustration on two data sets is provided.

Keywords Additive model · Event history · Logistic regression · Penalised likelihood

E. Elgmati
Department of Statistics, Tripoli University, Tripoli, Libya
e-mail: eelgmati@hotmail.com

R. L. Fiaccone
Department of Statistics, Universidade Federal da Bahia, Salvador, Brazil
e-mail: fiaccone@ufba.br

R. Henderson (✉) · J. N. S. Matthews
School of Mathematics and Statistics, Newcastle University, Newcastle, UK
e-mail: robin.henderson@ncl.ac.uk

J. N. S. Matthews
e-mail: john.matthews@ncl.ac.uk

1 Introduction

Assume that n individuals are observed at times $t = 1, 2, \ldots, \tau$, at each of which there is a binary event indicator Y_{it}, an at-risk indicator R_{it} and a p-dimensional predictable covariate vector x_{it}, which includes an intercept and may include dynamic covariates as considered by Aalen et al. (2004) and Fosen et al. (2006). The objective of interest in this work is the derivation of predictive event probabilities.

A variety of methods for dealing with longitudinal binary data of this form are available (Diggle et al. 2002). If τ is large however, it is sometimes convenient to borrow and adapt ideas developed for continuous time event history data. This was the approach taken by Borgan et al. (2007) in an analysis of data obtained as part of a programme of health education and sanitation improvements carried out in Salvador, Brazil. In their data set, daily records of diarrhoea were kept for 926 children over 455 days in what was termed Phase II of a larger programme, the event being the occurrence of diarrhoea (prevalence) or the start of an episode of diarrhoea (incidence). Analysis of similar data collected later in Phase III of the programme formed the original motivation for the work described in this paper, with emphasis now on prediction of events rather than modelling and estimation of covariate effects. We shall describe the data and our analyses in Sect. 6.

Henderson and Keiding (2005a, b) discussed prediction of single-event survival times and showed that accurate prediction should not be expected for even well-fitting statistical models with highly significant covariates. For recurrent event data there is perhaps hope for higher predictive power, at least dynamically, in the sense of learning about the subject under study as knowledge of their individual event frequency accrues. We are not aware of this having had close attention in the event-time literature, although there has been related work on the use of longitudinal biomarker information in dynamic/landmark prediction of single-event residual lifetimes (eg. Henderson et al. 2002; Proust-Lima and Taylor 2009; van Houwelingen and Putter 2011).

For this work we are interested in one-step-ahead prediction of events. Let $\mathcal{F}_t$ be the history or filtration generated by all covariates, observation patterns and event histories up to and including time t. Then define

$$\pi_{it} = \pi_{it}(\mathcal{F}_{t-1}) = P\left(Y_{it} = 1 \middle| \mathcal{F}_{t-1}\right)$$

together with a parametric model $\pi_{it}(\beta)$. Our purpose is to consider modelling and estimation strategies aimed at minimising the loss between observation and predictive probability, measured through some loss function $L(Y, \pi)$. We will concentrate on models incorporating a time-varying linear predictor

$$\alpha_{it} = \alpha_{it}(\beta) = \beta_{t0} + \beta_{t1}x_{it1} + \cdots + \beta_{tp}x_{itp}$$

with separate estimation at each time t and no smoothing of the coefficients $\{\beta_{tj}\}$ over time. Borgan et al. (2007) used the Aalen additive model

$$\pi_{it} = R_{it}\alpha_{it} \tag{1}$$

with least-squares estimation, an approach that is popular in the continuous-time event history literature (Martinussen and Scheike 2006). With binary data in mind, a perhaps more obvious approach would be based on logistic models

$$\pi_{it} = R_{it} \frac{e^{\alpha_{it}}}{1 + e^{\alpha_{it}}} = R_{it} \text{expit}(\alpha_{it}). \tag{2}$$

Either maximum likelihood or penalised maximum likelihood estimation as proposed by Firth (1993), Heinze and Schemper (2002) and Heinze (2006) can be used for estimation. In the following section we discuss the relative advantages and disadvantages of these methods as we see them. In Sect. 3 we argue for a very simple modification of the penalised likelihood approach. This modification is considered in detail for a simple special case in Sect. 4, before its use is illustrated in two applications in the next two sections. In Sect. 5 we consider data on timing of morphine requests for patients recovering from surgery, and in Sect. 6 we turn to the Phase III diarrhoea data referred to earlier.

2 Motivation

The Aalen additive modelling approach based on (1) has a large number of advantages. Parameters β_t can be estimated quickly and easily at each t using least squares:

$$\hat{\beta}_t = \left(X_t^T X_t \right)^{-1} X_t^T Y_t,$$

in the obvious vector/matrix form (and with dependence on $\{R_{it}\}$ supressed). Provided $X_t^T X_t$ is not singular, $\hat{\beta}_t$ always exists even when events are rare. If there are no events at time t then $\hat{\pi}_{it}$ is automatically zero, which is the nonparametric maximum likelihood estimator. Inference based on the cumulative coefficients

$$\hat{B}_{jt} = \sum_{u=1}^{t} \hat{\beta}_{ju}$$

is supported by powerful underpinning martingale theory, closed-form variance estimators are available and a martingale central limit theorem can usually be deployed. Inspection of plots of $\hat{B}_j(t)$ against t provides an extremely quick and effective tool to characterise covariate effects, especially if they change over time. A disadvantage of course is that $\hat{\pi}_{it}$ is not structurally bounded, and estimates bigger than one or less than zero can occur. When interest is in covariate effects this has no pragmatic consequence, and in our opinion the additive model is extremely effective when modelling event time data. For prediction however the picture is different, especially when events are rare, in which case $\hat{\pi}_{it}$ can often be negative and hence of at best limited use.

The logistic model (2) automatically bounds the estimates, at the cost of no closed-form estimation and the loss of the martingale machinery and simple interpretation of cumulative coefficients. Nonetheless, maximum likelihood estimation is straightforward and plots of $\hat{\beta}_{jt}$ against t are usually informative, provided some post-hoc smoothing is applied. On the other hand, when events are rare and $\sum_i Y_{it}$ is low, there can often be separation or quasi-separation (Albert and Anderson 1984), leading in principle to infinite parameter estimates and in practice non-convergence of iterative optimisation algorithms. Worryingly, and as pointed out by Heinze (2006), it can happen that likelihood convergence criteria are met and routines falsely indicate convergence. An illustration of this when the `glm` routine in R is used for estimation is presented in the Appendix. Obviously a careful analysis of a single data set would reveal separation problems, usually through highly extreme estimates and associated huge standard errors. In our case however we recall that we will repeat the logistic regression procedure for each t from 1 to τ, and if τ is large it is not feasible to pay close attention to each individual analysis. In the example of Sect. 5 for instance, τ =2,351.

A simple solution to separation problems was proposed in influential work of Heinze and Schemper (2002) and Heinze (2006). They advocated use of a modified score technique originally developed by Firth (1993) as a general approach to remove first-order bias in parametric problems. Firth showed that for exponential families with canonical parameterization, his method was equivalent to penalising the likelihood with a Jeffreys prior. Heinze and Schmemper pointed out that the same penalised likelihood technique is effective in overcoming separation problems in logistic regression. Thus, in our context, at each t, instead of maximising the log-likelihood $\ell(\beta_t)$, Heinze and Schemper (2002) propose maximisation of the penalised log-likelihood

$$\ell^*(\beta_t) = \ell(\beta_t) + \frac{1}{2}\log\left|I(\beta_t)\right| \tag{3}$$

where $I(\beta_t)$ is the information. As n increases the penalty term becomes negligible in comparison with $\ell(\beta_t)$, and the method reduces to standard logistic regression. Separation is not then an issue of course: it is a concern in the main for modest sample size n and relatively low event frequency. In these circumstances the method works well in stabilising regression coefficients and outperforms an alternative exact logistic regression approach (eg. Mehta and Patel 1995), with or without stratification and conditioning (Heinze 2006; Heinz and Puhr 2010), The procedure is fairly easy to apply and software is available for routine use. A price for predictive purposes is that in using a bias-correction technique for regression parameter estimates, the penalised likelihood approach introduces bias for predicted event probabilities. This will be illustrated in the next two sections.

First, we recap the discussion so far. We are interested for predictive purposes in fitting binary regression models at a potentially large number of timepoints τ, in circumstances where the event probability might be relatively low. The Aalen additive approach is pragmatic and successful for covariate effect analysis but weakened for predictive purpose by the presence of negative estimates. Standard logistic regression models can be unstable with few events and modest sample sizes. The Firth penalised approach is stable but in removing bias in regression coefficients it introduces bias into predicted event probabilities.

3 Logistic regression with modified penalised likelihood estimation

Our proposal is to use a logistic model with a modified penalised likelihood estimation approach, intermediate between the full Jeffreys penalty and no penalty at all. In exploring this suggestion, for the next two sections we will drop the subscript t and at-risk indicator and consider single samples of binary data.

We shall begin with a short historical diversion, which will involve a temporary change in notation and assumptions. Assume for now that there can be repeated design points, ie n_j observations at covariate value x_j, resulting in $Y_j \sim \mathrm{B}(n_j, \pi_j)$ events. The logistic model began to be seen as an alternative to probit regression in the late 1940s and early 1950s, advocated in particular by Berkson (eg. Berkson 1953 and references therein). Computational challenges meant that maximum likelihood was not generally feasible and hence there was close attention to alternative estimation procedures. One proposal was to look for a transformation of Y_j whose expectation is linear in x_j so that weighted least squares could be used. Anscome (1956) proposed the transformation

$$Z_j = \log\left(\frac{Y_j + \frac{1}{2}}{n_j - Y_j + \frac{1}{2}}\right)$$

and explained that the addition of 1/2 to numerator and denominator meant that Z_j was very nearly unbiased for $\log\big(\pi_j/(1-\pi_j)\big) = \beta x_j$. Cox (1970), quoting Anscome (1956) and also Haldane (1956), amplified the argument. Dropping the subscript j and starting with the "empirical logistic transform"

$$Z_\lambda = \log\left(\frac{Y+\lambda}{n-Y+\lambda}\right), \tag{4}$$

Cox let $U = (Y - n\pi)/\sqrt{n}$, so $E(U) = 0$ and $\mathrm{Var}(U) = \pi(1-\pi)$. Then

$$\begin{aligned} Z_\lambda - \log\left(\frac{\pi}{1-\pi}\right) &= \log\left\{1 + \frac{U}{\pi\sqrt{n}} + \frac{\lambda}{\pi n}\right\} - \log\left\{1 - \frac{U}{(1-\pi)\sqrt{n}} + \frac{\lambda}{(1-\pi)n}\right\} \\ &= \frac{U}{\pi(1-\pi)\sqrt{n}} + \frac{\lambda(1-2\pi)}{\pi(1-\pi)n} - \frac{(1-2\pi)U^2}{2\pi^2(1-\pi)^2 n} + o\left(\frac{1}{n}\right) \end{aligned}$$

and

$$E[Z_\lambda] - \log\left(\frac{\pi}{1-\pi}\right) = \frac{(1-2\pi)(\lambda - \frac{1}{2})}{\pi(1-\pi)n} + o\left(\frac{1}{n}\right). \tag{5}$$

Setting $\lambda = \frac{1}{2}$ removes the first order bias term, as claimed by Anscome (1956). With characteristic prescience, Cox wrote "It is interesting, but not essential for the argument, that we can nullify the term in $1/n$ in [our (5)] by a single choice of [our λ], independent of [our π]", thus pre-empting, for this special case, the more general argument of Firth.

As an illustration of his bias correction technique, Firth (1993) also used the single binomial observation model. The target parameter is $\beta = \log\{\pi(1-\pi)\}$, the information is proportional to $\pi(1-\pi)$ and the penalised log-likelihood is simply

$$\ell^*(\beta) = Y\log\pi + (n-Y)\log(1-\pi) + \frac{1}{2}\log\{\pi(1-\pi)\}. \tag{6}$$

Maximisation of $\ell^*(\beta^*)$ leads to

$$\beta^* = \log\left(\frac{Y+\frac{1}{2}}{n-Y+\frac{1}{2}}\right),$$

which is precisely the Anscome empirical logistic transformation.

If, instead, we choose to maximise the alternatively penalised log-likelihood,

$$\ell^*(\beta) = Y\log\pi + (n-Y)\log(1-\pi) + \lambda\log\{\pi(1-\pi)\}, \tag{7}$$

then

$$\beta^* = \log\left(\frac{Y+\lambda}{n-Y+\lambda}\right) \tag{8}$$

is obtained, ie the starting point (4) for the Cox (1970) derivation. By varying λ from 0 to 1/2 we can move between standard logistic regresssion (almost unbiased for π but biased for β, undefined for $Y=0$ or $Y=n$) to the Firth-corrected version (almost unbiased for β but biased for π, well-defined at $Y=0$ and $Y=n$).

We now return to the general case and state our proposal. When interest is in prediction using a logistic model but events are rare, then a modified penalised likelihood model might be advocated: estimate by maximisation of

$$\ell^*(\beta) = \ell(\beta) + \lambda\log\left|I(\beta)\right|. \tag{9}$$

for a given λ. Typically we would take λ to be between 0 and 0.5.

3.1 Bias in regression parameter β

Suppose β is the parameter interest. Let $U(\beta)$ be the score and $I(\beta)$ be the observed information. Let the first-order bias of the maximum likelihood estimator $\hat{\beta}$ be $b(\beta)/n$.

Firth (1993) considered a modified estimating equation of the form

$$U^*(\beta) = U(\beta) + A(\beta),$$

where $A(\beta)$ is $O(1)$ as sample size n increases. Firth showed that the bias of the estimator solving $U^*(\beta)=0$ is

$$\frac{b(\beta)}{n} + i^{-1}(\beta)\alpha(\beta) + O\left(n^{-3/2}\right) \tag{10}$$

where $i(\beta)$ is the expected information and $\alpha(\beta)$ is the null expectation of $A(\beta)$. He also showed that if β is the canonical parameter in an exponential family model, then choosing $A(\beta)$ such that

$$\alpha(\beta) = \frac{1}{2} \frac{\partial \log \left| i(\beta) \right|}{\partial \beta}$$

removes the first-order bias term. Hence from (10)

$$b(\beta) = -\frac{n}{2} i^{-1}(\beta) \frac{\partial \log \left| i(\beta) \right|}{\partial \beta}. \tag{11}$$

If we set

$$U^*(\beta) = U(\beta) + \lambda \frac{\partial \log \left| i(\beta) \right|}{\partial \beta},$$

then we can estimate by solving $U^*(\beta) = 0$ using a simple adjustment to the Newton–Raphson estimation procedure for the Firth penalty, as described by Heinze and Schemper (2002). The new modified estimator β^* has from (10) and (11) bias

$$\left(\lambda - \frac{1}{2} \right) i^{-1}(\beta) \frac{\partial \log \left| i(\beta) \right|}{\partial \beta} + O(n^{-3/2}). \tag{12}$$

In the logistic case with

$$\pi_i = \pi_i(\beta) = \text{expit}(\beta x_i),$$

the observed information

$$I(\beta) = \sum_i \pi_i (1 - \pi_i) x_i x_i^T$$

does not depend on responses $\{Y_i\}$ and is hence equivalent to the expected information, conditional on covariates.

3.2 Bias in predictive probability π

Now let $\pi_0(\beta)$ be the estimated event probability of a new observation with covariate x_0. In order to study the bias of $\pi_0(\beta^*)$ we write

$$\beta^* = \beta + B_n$$

where B_n is a sample-dependent random variable with mean $\mu_n + O(n^{-3/2})$ and variance $\Sigma_n + O(n^{-3/2})$, where

$$\mu_n = \left(\lambda - \frac{1}{2} \right) i^{-1}(\beta) \frac{\partial \log \left| i(\beta) \right|}{\partial \beta}$$

is the leading bias term in (12) and

$$\Sigma_n = i(\beta)^{-1}$$

is the leading order term in the (robust) variance of β^*. Then recalling that we are interested in small $e^{\beta^* x_0}$, we can take the expansion

$$\frac{e^{\beta^* x_0}}{1 + e^{\beta^* x_0}} = \sum_{k=1}^{\infty} (-1)^{k+1} e^{k\beta^* x_0}. \tag{13}$$

Since β^* differs from the MLE by terms that are $O(n^{-1})$, we proceed by assuming that β^* is Normally distributed and hence that the expectation of (13) can be approximated by the finite sum

$$E\left[\pi_0(\beta^*)\right] \simeq \sum_{k=1}^{K} (-1)^{k+1} \exp\left\{k(\beta + \mu_n)x_0 + k^2 x_0^T \Sigma_n x_0/2\right\}. \tag{14}$$

Our experience from simulations and the exact calculations of the following section, is that good approximations are obtained if K is two or more. For example, Table 1 provides estimates for three new observations when there is a single standard Normal covariate and a sample of size $n = 100$ is used for estimation of β. Given that λ is positive, there is overestimation as expected, which becomes worse as λ approaches 0.5. The theoretical values of expected $\pi_0^*(\beta^*)$ from (14) match the empirical means from 3,000 simulated samples very closely for $K = 2$ and $K = 3$ but are too high at $K = 1$.

4 Two groups

Examination of the simple two-group situation is informative as exact calculation is possible. We will once more change notation temporarily. Let Y_0 be the number of events amongst n_0 subjects in group zero with $x = 0$, and Y_1 and n_1 be the corresponding values for group one with $x = 1$. The event probabilities are π_0 and π_1, parametrised in a logistic model as

$$\pi_0 = \text{expit}(\beta_0) \qquad \pi_1 = \text{expit}(\beta_0 + \beta_1).$$

The maximum modified penalised likelihood estimates are

$$\beta_0^* = \log\left(\frac{Y_0 + \lambda}{n_0 - Y_0 + \lambda}\right) \qquad \beta_1^* = \log\left(\frac{Y_1 + \lambda}{n_1 - Y_1 + \lambda}\right) - \log\left(\frac{Y_0 + \lambda}{n_0 - Y_0 + \lambda}\right)$$

and

$$\pi_0^* = \frac{Y_0 + \lambda}{n_0 + 2\lambda} \qquad \pi_1^* = \frac{Y_1 + \lambda}{n_1 + 2\lambda}.$$

Table 1 Predictive probabilites with modified penalised likelihood estimation

	$x_0 = (1, -1)$	$x_0 = (1, 0)$	$x_0 = (1, 1)$
		True π	
	0.029	0.047	0.076
Empirical mean $\pi_0^*(\beta^*)$ from 3,000 simulations			
$\lambda = 0.1$	0.032	0.046	0.075
0.2	0.034	0.047	0.077
0.3	0.035	0.049	0.079
0.4	0.036	0.050	0.081
0.5	0.037	0.052	0.083
$E\left[\pi_0(\beta^*)\right]$ from (14) with $K = 1$			
$\lambda = 0.1$	0.035	0.048	0.083
0.2	0.037	0.050	0.085
0.3	0.038	0.052	0.088
0.4	0.040	0.054	0.091
0.5	0.041	0.056	0.094
$E\left[\pi_0(\beta^*)\right]$ from (14) with $K = 2$			
$\lambda = 0.1$	0.033	0.045	0.074
0.2	0.034	0.047	0.076
0.3	0.035	0.049	0.078
0.4	0.037	0.050	0.080
0.5	0.038	0.052	0.082
$E\left[\pi_0(\beta^*)\right]$ from (14) with $K = 3$			
$\lambda = 0.1$	0.033	0.046	0.075
0.2	0.034	0.047	0.077
0.3	0.036	0.049	0.080
0.4	0.037	0.051	0.082
0.5	0.039	0.053	0.084

Sample size $n = 100$, $\beta = c(-3, -1)$, one standard Normal covariate

The bias and mean square error of each of π_0^* and π_1^* can easily be found. For example, π_0^* has bias

$$\mathrm{b}(\pi_0^*) = \frac{\lambda(1 - 2\pi_0)}{n_0 + 2\lambda}$$

and mean square error

$$\mathrm{MSE}(\pi_0^*) = \frac{n_0\pi_0(1 - \pi_0) + \lambda^2(1 - 2\pi_0)^2}{(n_0 + 2\lambda)^2}.$$

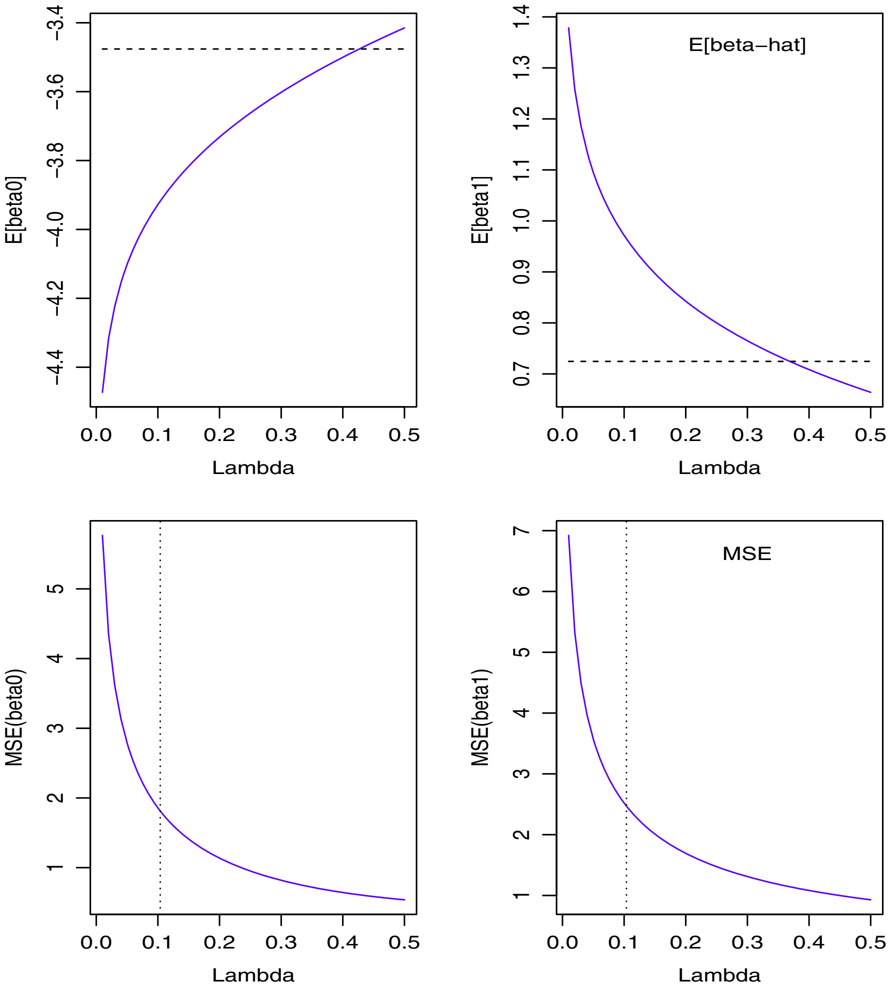

Fig. 1 Expected values and mean square errors of regression parameters β^* with modified penalised likelihood estimation for the two-group case. Sample sizes $n_0 = n_1 = 50$ and event probabilities $\pi_0 = 0.03$ and $\pi_1 = 0.06$. The *horizontal dashed lines* in the *upper* plots show the true values, and the *vertical dotted lines* in the *lower* plots indicate the λ that gives the minimum combined mean square error of prediction

As an aside we note that if interest was only in π_0 then this mean square error could be minimised at

$$\lambda^{\text{opt}} = 2\pi_0(1-\pi_0)/(1-2\pi_0)^2, \tag{15}$$

showing that the modified technique can outperform both standard maximum likelihood and the Firth/Jeffreys penalty approach.

Properties of β_0^* and β_1^* need to be obtained numerically, by averaging over the binomial distributions of Y_0 and Y_1. Figures 1 and 2 show the expected values and mean square errors of the modified penalised likelihood estimators β^* and π^* respectively, at $n_0 = n_1 = 50$ and $\pi_0 = 0.03$, $\pi_1 = 0.06$. As expected, each element of the

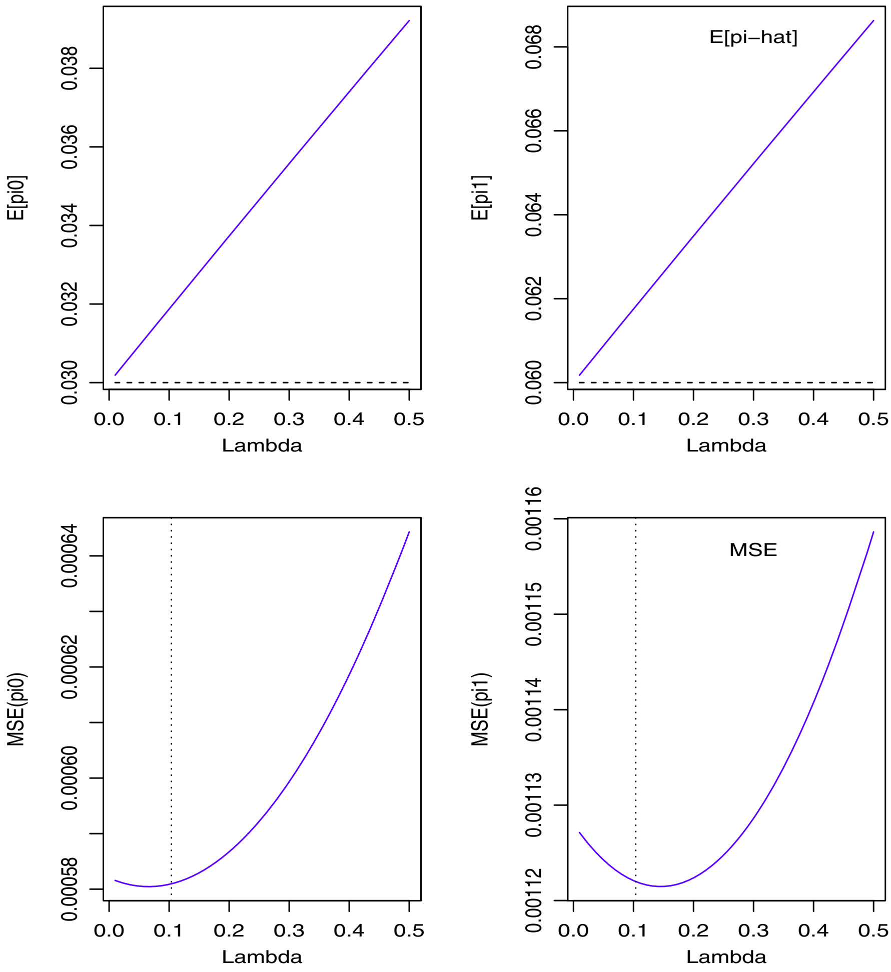

Fig. 2 Expected values and mean square errors of predictive probabilities π_λ^* with modified penalised likelihood estimation for the two-group case. Sample sizes $n_0 = n_1 = 50$ and event probabilities $\pi_0 = 0.03$ and $\pi_1 = 0.06$. The *horizontal dashed lines* in the *upper* plots show the true values, and the *vertical dotted lines* in the *lower* plots indicate the λ that gives the minimum combined mean square error of prediction

regression parameter β^* is almost unbiased for $\lambda = 0.5$ but quite severely biased at λ near zero. The predictive probabilities π^* have the opposite properties, as expected. Both β^* and π^* have decreasing variance as λ increases (not shown). Combining this with decreasing bias obviously leads to decreasing mean square error for the regression parameter β^*. For π^* by contrast, the bias increases with λ, leading in this example to a local minimum in root mean square error for both probability estimates.

Turning to predicting a new observation, we might consider the value of λ that minimises an overall mean square error

$$w_0 E\left[\left(\pi_0^* - \pi_0\right)^2\right] + (1 - w_0) E\left[\left(\pi_1^* - \pi_1\right)^2\right]$$

where w_0 reflects the weight to be attached to $x = 0$. The optimal value of λ solves a cubic equation with coefficients depending on π_0, π_1, n_0, n_1 and w_0, and no simple expression is available, although it is easy to calculate numerically. When $n_0 = n_1$ there is a closed form, namely

$$\frac{2w_0\pi_0(1-\pi_0) + 2(1-w_0)\pi_1(1-\pi_1)}{w_0(1-2\pi_0)^2 + (1-w_0)(1-2\pi_1)^2}. \tag{16}$$

This value is indicated in Figs. 1 and 2, evaluated at $w_0 = 0.5$. It is interesting to note that its value is extremely close in this example to λ^{opt} calculated from (15) but with π_0 replaced by the marginal event probability, $(\pi_0 + \pi_1)/2$, namely $\lambda^{\text{opt}} = 0.1038$ from (15) compared with $\lambda = 0.1031$ from (16).

5 Application I: patient controlled analgesia

We consider event time data from $n = 65$ patients monitored for 2 days following stomach surgery. The event is the self-administration of a bolus of painkiller (morphine), and the timescale is minutes. There are two groups: the morphine bolus was set at 2 mg for 39 patients (group 0) and the pump then locked out automatically for 8 min; for 35 patients (group 1) the bolus was 1mg and the lockout time was 4 min. Covariate information is available on operation type (extensive incision = 1, other = 0), gender, categorised age, weight and initial loading of painkiller. We also defined a dynamic covariate D_{it} as the event rate over the previous 3 h for patient i, with proper allowance for lockout period and with predictions not starting until 180 min so that D_{it} is always well defined. Following Aalen et al. (2004), Fosen et al. (2006) and Borgan et al. (2007), rather than using D_{it} directly in modelling, we replaced it with the residual from a linear model of D_{it} on x_i, the remaining covariates. In this way the indirect effects of x_i on D_{it} do not dilute the estimated direct effects of x_i on event rates.

Over the 2,880 min 2-day period, there were 529 timepoints on which no events occured. These were taken out of the analyses, leaving $\tau = 2,351$ event times. Incidence at these varied from 1.6 to 21.7 % events, with marginal rate 3.6 %. Excluding the first 180 min, logistic regression successfully converged at just 104 time-points, the R routine `glm` indicated non-convergence at 1,557 time-points, and there were 525 points where there was clear false convergence, ie the routine indicated convergence but at least one standard error was over 1,000. A logistic modelling approach with standard maximum likelihood is clearly not appropriate for these data.

Useful information and valid inferences can be obtained quickly and easily however using the Aalen additive model for discrete time and with dynamic covariates, as described by Borgan et al. (2007). Figure 3 shows the cumulative estimates $\hat{B}_t$ obtained from the fit. Patients in group 1, with the lower bolus size, had more events than those in group 0, as expected. Operation type, gender, age and initial loading are all important in determining event rate, but the most striking feature of Fig. 3 is the strong and hugely significant effect of the dynamic covariate, essentially recent event rate. This is an indicator of heterogeneity between patients, akin to a frailty effect.

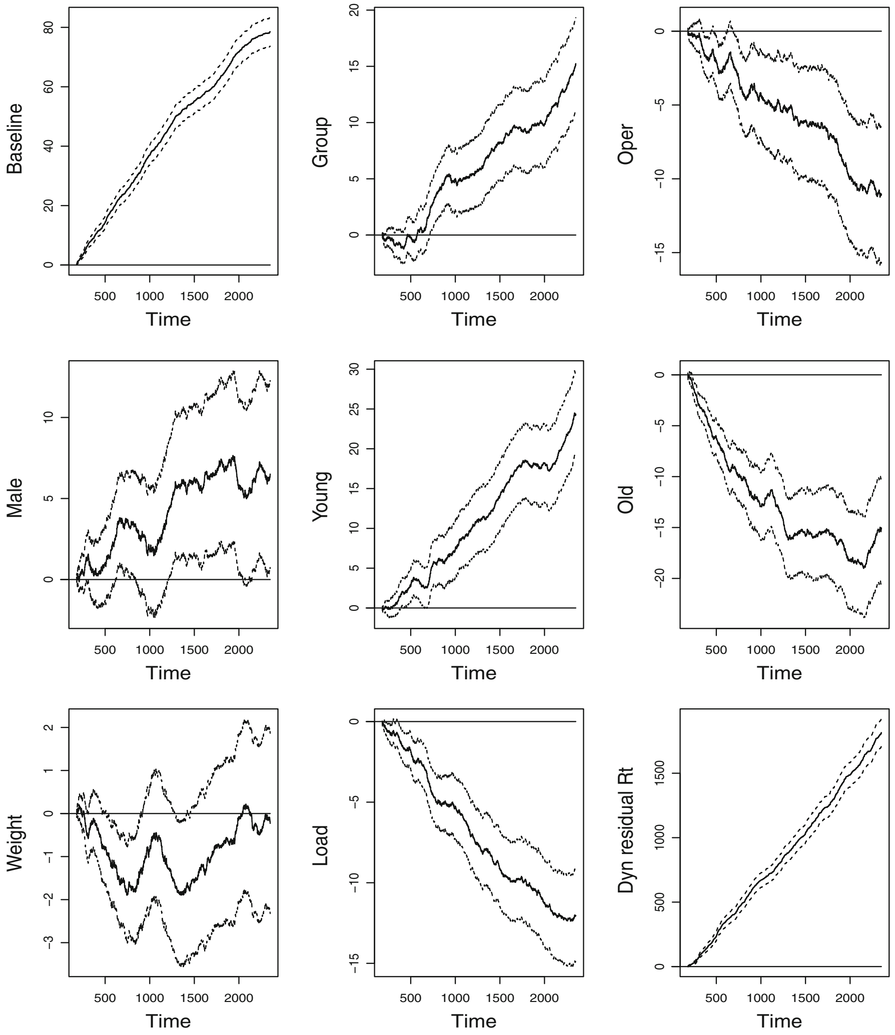

Fig. 3 PCA data. Cumulative regression coefficients $\hat{B}_t$ for Aalen additive model fit, with approximate 95 % pointwise confidence intervals

Turning to our main focus, prediction, in the following we use leave-one-out methods so that predictive probabilty estimates $\pi^*_{(i)t} = P(Y_{it} = 1 \big| \beta^*_{(i)}, \mathcal{F}_{t-1})$ are based on parameters estimated with the patient of interest excluded. Individual-specific histories up to time t are incorporated in $\mathcal{F}_{t-1}$ so that the dynamic covariate D_{it} can be used for prediction. Of a total of just over 150,000 patient-time estimates $\pi^*_{(i)t}$, some 31 % were negative when the Aalen additive model was used for estimation. Hence we do not consider this model for predictive purposes.

The Firth and modified penalised likelihood methods always converged and always provided what seemed to be reasonable predictive probabilities. Table 1 summarises the results for a variety of penalty parameters λ. The table gives the Brier score (BS), the summed jackknife deviance residuals

$$\mathrm{JD} = -\sum_t \sum_i \left\{ Y_{it} \log \pi^*_{(i)t} + (1 - Y_{it}) \log \left(1 - \pi^*_{(i)t} \right) \right\} \tag{17}$$

and the predictive total number of events

$$\mathrm{PT} = \sum_t \sum_i \pi^*_{(i)t}. \tag{18}$$

The Brier score is a poor measure for rare events as better scores can sometimes be obtained from, for example, a constant prediction of even zero (eg. Ferro and Stephenson 2012; Jachan et al. 2009). It is included for completeness. Both the Brier score and the preferred jackknife deviance are minimised in Table 1 at $\lambda = 0.1$ within the set of values considered. Changes in deviance are highly significant. The predictive totals can be compared with the observed post-burn-in total of 4,373 events. As expected, the Firth method ($\lambda = 0.5$) leads to over-prediction, though it is more severe than might have been anticipated. The over-prediction is reduced as λ is decreased and event probability estimates become less biased. Overall it seems that the modified penalised likelihood method with $\lambda = 0.1$ finds a nice compromise for these data between convergence problems with standard maximum likelihood and biased probability estimates with the Firth/Jeffreys penalty.

Figure 4 shows the cumulative total observed and leave-one-out predicted numbers of events for two subjects, one with a high number of events and one with a low number. They were chosen as the people closest to the 90 and 10 % points of the ordered total event count data. Predictions were made using both the Firth penalty and the modified penalty at $\lambda = 0.1$. The former method severely over-predicts events, but the latter works well for these people.

6 Application II: infant diarrhoea

Our second application is to data on occurrence of infant diarrhoea. As part of Phase III of the Blue Bay sanitation programme in Salvador, Brazil, daily diarroea records were kept for n =1,127 infants over $\tau = 227$ days between October 2003 and May 2004. Similar data collected in Phase II of the programme were analysed in Borgan et al. (2007). Events correspond to the 1st day of a new episode of diarrhoea, with episodes considered to end when 3 or more days occurred without diarrhoea. We consider five baseline covariates and one dynamic one. The baseline covariates are mothers' age, housing occupation density (1 if two or more people per room, 0 otherwise), presence of nearby open sewerage, waste collection frequency (0 if frequent, 1 if less frequent) and whether the local streets had been paved. The dynamic covariate D_{it} is the rate of previous events before time t, defined as number of events divided by number of days at risk. We took a 10-day burn-in and all summaries below exclude this period.

The marginal incidence rate was 0.8 %, with daily rates varying from 0.1 to 4.6 %, excluding 8 days on which no events occur. Although the sample size is large, the

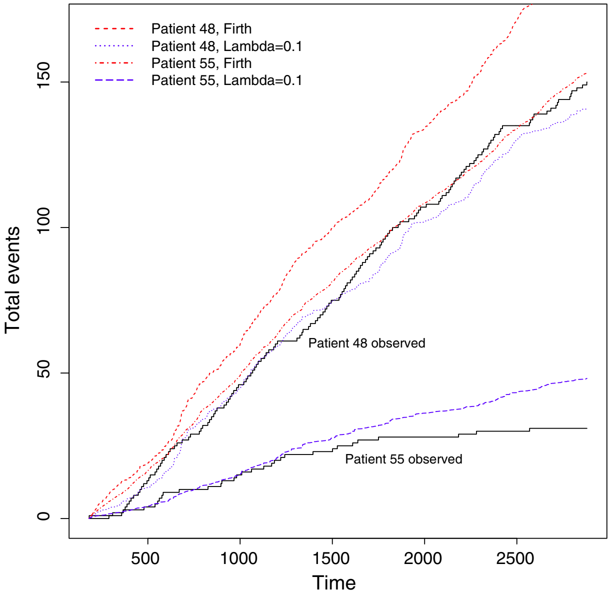

Fig. 4 PCA data. Observed and predicted cumulative number of events for two patients

low marginal incidence rate means that separation and convergence problems occur. When attempting to fit a logistic model to the daily data, standard maximum likelihood converged just 86 times, and there was false convergence (likelihood convergence but at least one standard error above 1,000) for 123 days. The R routine `glm` indicated convergence for all 209 analyses.

Figure 5 shows the cumulative regression coefficients when an Aalen additive model is fit to the data using the same methods as used by Borgan et al. (2007) in their analysis of data from the earlier Phase II of the Blue Bay programme. Incidence is lower for infants with relatively old mothers, but is higher for infants living in relatively crowded accommodation or areas with poor garbage services. Presence of open sewerage and unpaved roads had little effect on incidence, but again the dynamic covariate is highly significant, with compelling evidence that infants with a history of diarrhoea are more prone to future occurrence. Turning to prediction, 24 % of the approximately 250,000 infant-days had negative predictive probabilities under this model.

Table 3 summarises predictive performance when we use Firth or modified penalised likelihood to fit a logistic model. The observed total number of events was 1,393, and with that in mind the picture is very similar to that of Table 2. Choosing to penalise at a value of $\lambda = 0.1$ or 0.2 is to be recommended.

7 Discussion

We have proposed a modified penalised likelihood for logistic regression when interest is in prediction. Choosing weight λ around 0.1 seems to give a good trade-off

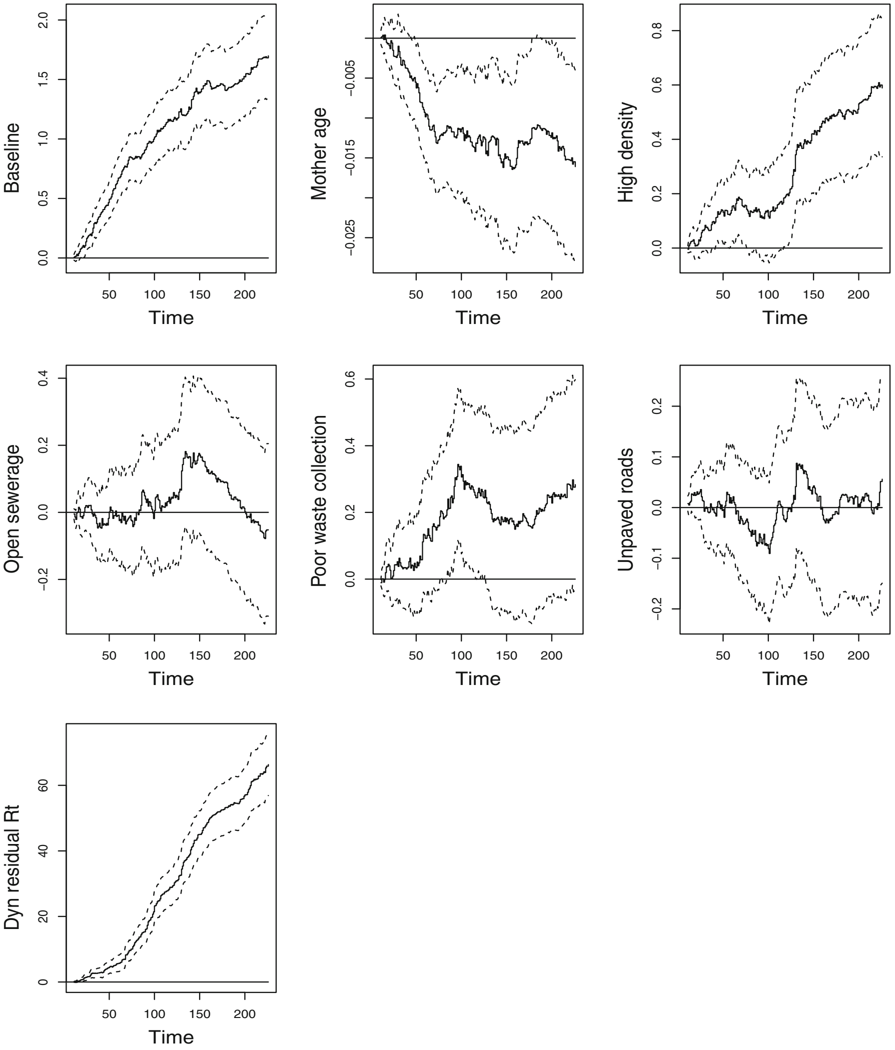

Fig. 5 Diarrhoea data. Cumulative regression coefficients $\hat{B}_t$ for Aalen additive model fit, with approximate 95 % pointwise confidence intervals

between separation and stability issues when events are rare, and the bias in probability estimates π that is introduced when attempting to remove first-order bias in estimates of regression parameters β. Although our motivation in this work has been on dynamic analysis of recurrent event data, the proposed technique might be useful for standard single-sample analysis of binary data, and perhaps in more general parametric modelling. A fuller careful investigation is needed.

When there are recurrent events, another method that can deal with separation issues is to undertake some form of smoothing, perhaps to analyse data pooled over a moving window of time points rather than separately at each. Missing data need

Table 2 PCA data

λ	BS	JD	PT
0.5	0.0334	18031.6	11106.5
0.4	0.0321	16768.4	9720.1
0.3	0.0310	15578.4	8298.3
0.2	0.0301	14585.4	6850.6
0.1	0.0298	14190.2	5387.7
0.05	0.0300	14735.6	4653.6

Predictive probability summary for logistic models with modified penalised likelihood estimation: BS = Brier score; JD = jackknife deviance statistic (17) PT = predictive total (18)

Table 3 Diarrhoea data

λ	BS	JD	PT
0.5	0.00783	7603.1	2081.919
0.4	0.00781	7562.0	1948.586
0.3	0.00779	7531.0	1814.107
0.2	0.00777	7515.4	1677.859
0.1	0.00776	7528.5	1540.032
0.05	0.00776	7547.4	1470.508

Predictive probability summary for logistic models with modified penalised likelihood estimation: BS = Brier score; JD = jackknife deviance statistic (17); PT = predictive total (18)

particularly careful attention in this case, as individuals may be observed in some but not all times in the window. Missing data can also be problematic if the purpose is to predict events more than one time point ahead. Assessing predictive accuracy then becomes challenging, as does inclusion of dynamic covariates and some form or marginalisation is likely to be required. For example we might be interested in predicting at time t the cumulative number of events over the period $t+1$ to $t+k$. Our preferred model might assume that the probability of an event at any time can depend on the occurrence or not of an event at the immediately preceding time. Unless $k=1$ this data is not available and either we marginalise over possible values or we restrict ourselves to dynamic covariates that exclude the $k-1$ most recent time points. Another issue that needs to be dealt with is when the occurrence of an event immmediately precludes further events for a time. In the patient controlled analgesia application for instance, the analgesia machinery automatically locked out for either 4 or 8 min following each dose. In the diarrhoea application, a new episode of diarrhoea cannot occur unless there have been at least two diarrhoea-free days since the last episode. Episodes being of random length (mostly 1–5 days) brings another difficulty. This is the focus of current work by the authors.

Acknowledgments The research of Rosemeire Fiaccone was supported in part by National Council of Technological and Scientific Development—CNPq, Brazil (Num. 237094/2012-6, 480614/2011-3). Robin Henderson benefited from participation in Deutsche Forschungsgemeinschaft research programme FR 3070/1-1. We are grateful for the comments of the reviewers.

Appendix: logistic regression with separation not detected in R

In the following, `y` is a vector of length 100, with all elements zero except the first, which is one, and `x1` is a vector of 50 zeros followed by 50 ones, representing two equally sized groups. If we attempt to fit the logistic regression

$$\pi_x = P\big(Y = 1\big|x\big) = \text{expit}\big(\beta_0 + \beta_1 x\big)$$

then clearly a perfect fit is obtained at $\hat{\beta}_0 = \text{logit}(1/50) = -3.892$ and $\hat{\beta}_1 = -\infty$. Some R (version 3.1.2) output, edited to remove unnecessary material (marked by `[...]`, is:

```
> fit=glm(y˜x1,family=’binomial’)
> summary(fit)

[...]

             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.892      1.010  -3.853 0.000117 ***
x1            -17.674   4134.091  -0.004 0.996589

[...]

> unique(fit$fitted.values)
[1] 2.000000e-02 4.305023e-10
> fit$converged
[1] TRUE
```

Of most concern is the statement of convergence, which is true because the maximised likelihood has indeed converged: moving either of the coefficients away from their current values leads to no improvement. The fitted probabilities $\hat{\pi}_0$ and $\hat{\pi}_1$ are accurate but clearly $\hat{\beta}_1$ is unrealistic. Uncritical assessment of the results might lead to this problem being missed.

If we use the Firth correction as implemented in Kosmidis’ bias reduction package `brglm`, we obtain:

```
> fitbr=brglm(y˜x1,family=’binomial’)
> summary(fitbr)
[...]

             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.497      0.837  -4.177 2.95e-05 ***
x1             -1.119      1.662  -0.673    0.501

> unique(fitbr$fitted.values)
[1] 0.029411765 0.009803922
> fitbr$converged
[1] TRUE
```

Hence the coefficients are stabilised, at the expense of higher values of $\hat{\pi}_0$ and $\hat{\pi}_1$ as expected. Heinze' package `logistf` gives the same results.

References

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

Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71:1–10

Anscome FJ (1956) On estimating binomial response relations. Biometrika 43:461–464

Berkson J (1953) A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. J Am Statist Assoc 48:565–599

Borgan Ø, Fiaccone RL, Henderson R, Barreto ML (2007) Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. Scandinavian J Statist 34:53–69

Cox DR (1970) Analysis of binary data, 1st edn. Chapman and Hall, London

Diggle PJ, Heagerty PJ, Liang K-Y, Zeger S (2002) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford

Ferro CAT, Stephenson DB (2012) Deterministic forecasts of extreme events and warnings. In: Jolliffe IB, Stephenson DB (eds) Forecast verification: a practitioner's guide in atmospheric science, 2nd edn. Wiley, Chichester

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27–38

Fosen J, Borgan Ø, Weedon-Fekær H, Aalen OO (2006) Dynamic analysis of recurrent event data using the additive hazard model. Biometr J 48:381–398

Haldane JBS (1956) The estimation and significance of the logarithm of a ratio of frequencies. Ann Human Genet 20:309–311

Heinz G, Puhr R (2010) Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. Statist Med 29:770–777

Heinze G, Schemper M (2002) A solution to the problem of separation in logistic regression. Statist Med 21:2409–2419

Heinze G (2006) A comparative investigation of methods for logistic regression with separated or nearly separated data. Statist Med 25:4216–4226

Henderson R, Diggle PJ, Dobson A (2002) Identification and efficacy of longitudinal markers for survival. Biostatistics 3:33–50

Henderson R, Keiding N (2005a) Individual survival time prediction using statistical models. (Forudsigelse af individuelle levetider ved hjaelp af statistuiske modeller). Danish Med J 167/10:1174–1177

Henderson R, Keiding N (2005b) Individual survival time prediction using statistical models. J Med Ethics 31:703–706

Jachan M, Feldwisch H, Posdziech F, Brandt A, Altenmüller D-M, Schulze-Bonhage A, Timmer J, Schelter B (2009) Probabilistic forecasts of epileptic seizures and evaluation by the Brier score. Fourth Eur Conf Int Federation Medi Biol Eng Proc 22:1701–1705

Martinussen T, Scheike TH (2006) Dynamic regression models for survival data. Springer, New York

Mehta CR, Patel NR (1995) Exact logistic regression: theory and examples. Statist Med 14:2143–2160

Proust-Lima C, Taylor JMG (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics 10:535–549

van Houwelingen H, Putter H (2011) Dynamic prediction in clinical survival analysis. Chapman and Hall/CRC Press, London