

Analysis of Interval-Censored Longitudinal Data with Application to Onco-Haematology

*Analisi di Dati Longitudinali Soggetti a Censura Intervallare
con Applicazione in Onco-Ematologia*

Paola De Lorenzo^{§*}, Robin Henderson[°] and Maria Grazia Valsecchi^{*}

[§] Istituto di Statistica Medica e Biometria, Università di Milano, Milano.

^{*} Dipartimento di Medicina Clinica, Prevenzione e Biotechnologie Sanitarie,
Università di Milano-Bicocca, Via Cadore, 48 20052 – Monza (MI).

[°] Mathematics and Statistics, Lancaster University, Lancaster, UK.

paola.delorenzo@unimib.it

Riassunto: L'obiettivo di questo lavoro è lo studio delle misurazioni ripetute soggette a censura intervallare. In particolare, si considerano qui quelle rilevate nell'ambito di studi clinici. Lo scopo è valutarne la relazione con il tempo all'evento, dove per evento si intende una misurazione oltre una soglia prefissata. L'uso del modello di Cox in questo ambito è controverso per vari motivi. Si è pertanto esplorata la possibilità di applicare i metodi classici per l'analisi di dati longitudinali. Il modello che si propone è un loro adattamento che consente un'appropriata descrizione dei dati, mantenendo nel contempo limitata la complessità dei calcoli e agevole l'interpretazione dei risultati.

Keywords: Censored Data, Latent Variables, Repeated Measurements, Survival Analysis

1. Introduction

The analysis of repeated measurements on a biomarker, either alone or jointly with the analysis of the time to the event of interest, is an area of active research. Nevertheless, we are not yet able to deal in complete generality with these complex data, which frequently consists of error-prone, sparse and intermittent values. In many cancer studies, they arise in the framework of clinical trials and thus their relationship with prognosis is a primary focus. In such a setting, the Cox model is regarded as the standard technique for the analysis, but crucial issues may discourage its application:

- a. A biomarker is almost always an internal time-dependent variable, in that it depends on the disease process [1].
- b. The actual knowledge of the biomarker's trajectory over time is unlikely to meet the requirements of the Cox model, which would need the updates at each event time, for each subject still at risk.
- c. Somewhat intermediate between point (a) and (b) is the problem of informative drop-out, that may occur when scheduled measurements are not taken.
- d. The estimate of the impact of the up-dated biomarker on the hazard is constant as a consequence of the proportional hazards assumption and does not keep track of the individual trajectory of the variable.

The aim of this work is to illustrate an alternative approach to the analysis of studies in which the event occurs when the biomarker measurement passes a certain threshold, complicated by interval-censoring of the biomarker values. The work is motivated by a study in the childhood leukaemia in which the event, i.e. the relapse, is defined as the point at which an interval-censored measure of leukemic cells passes a threshold.

We propose a linear mixed model with a Gaussian stochastic process that can be used to track the biomarker trajectory and to estimate the probability of event occurrence. This model for interval-censored data has been developed within the classic approach to longitudinal data analysis, that was previously adapted to allow for left-censored data, only [2].

2. Methods

If $Y_i(t_i)$ denotes the vector of the longitudinal biomarker's measurements on the i -th subject, $i=1, \dots, N$ taken at times $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$, then a linear mixed model for Gaussian responses may be written as [3]

$$Y_i(t_i) = X_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \omega_i(\mathbf{t}_i) + \boldsymbol{\varepsilon},$$

where $X_i\boldsymbol{\beta}$ is the mean response described by a linear function of r (possibly time-dependent) covariates through r parameters $\boldsymbol{\beta}$; $\mathbf{Z}_i$ is a $n_i \times p$ matrix of covariates that may depend on time and are related to the p -dimensional vector of random effects $\boldsymbol{\gamma}_i$. These are assumed to be normal distributed with mean zero and covariance matrix $\mathbf{B}$ and independent from $\omega_i(\mathbf{t}_i)$, a Gaussian process with mean zero and covariance matrix $\boldsymbol{\Sigma}$. The model is completed by the error term $\boldsymbol{\varepsilon}$ which is also assumed to be normal with null mean and covariance $\sigma_\varepsilon^2 \mathbf{I}$.

The hypothesis of independence among all of the random components eventually leads to

$$\mathbf{Y}_i \sim N(X_i\boldsymbol{\beta}, V_i),$$

where

$$V_i = \mathbf{Z}_i \mathbf{B} \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i + \sigma_\varepsilon^2 \mathbf{I}_i.$$

In the biological field, it is reasonable to hypothesise that different individuals be characterized both by different fluctuations around the common, overall mean (random intercept) and by different profiles in time (random slope). We thus take $\mathbf{Z}_i = [\mathbf{1}, \mathbf{t}_i]$ where each element of the vector $\mathbf{t}_i$ is the number of days from diagnosis until the date at which the corresponding biomarker measurement has been taken. Moreover, $\boldsymbol{\gamma}_i = [\gamma_{i1}, \gamma_{i2}]$, $\boldsymbol{\gamma}_i \sim N(\boldsymbol{\theta}, \mathbf{B})$, $\mathbf{B} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$. We further assume that the serial correlation

term $\omega_i(\mathbf{t}_i)$, is such that its components ω_{ij} $j=1, \dots, n_i$ follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_i$, whose elements are $Var(\omega_{ij}) = \sigma_\omega^2$ and $Cov(\omega_{ij}, \omega_{ik}) = \sigma_\omega^2 \exp\{-\alpha|t_{ij}-t_{ik}|\}$, where $\alpha > 0$, for any $k \neq j$. Note that, the more distant the serial observations, the weaker their correlation link.

Due to the limitations of the measurement procedure, the biomarker's data $\mathbf{Y}_i$ may be interval-censored, in that the available data are the extremes of the interval $[a_{ij}, b_{ij})$ in which the “true” underlying, continuous measure $\mathbf{Y}_i^*$ lies. Then, the log-likelihood function, given also the assumptions above, takes the form

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^N \log \int_{a_{i1}}^{b_{i1}} \dots \int_{a_{in_i}}^{b_{in_i}} \frac{1}{(2\pi)^{n_i/2} |\mathbf{V}_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{u} - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1} (\mathbf{u} - \boldsymbol{\mu}_i) \right\} du_1, \dots, du_{n_i}$$

where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma_1, \sigma_2, \rho, \alpha, \sigma_\omega, \sigma_\varepsilon]$ is the vector of parameters and $\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$ is the mean level. The estimate $\hat{\boldsymbol{\theta}}$ can be computed through direct maximization of the likelihood. We applied the Nelder-Mead algorithm, which does not rely on the gradient, and fitted a quadratic surface in $\hat{\boldsymbol{\theta}}$ to compute $\text{Var}(\hat{\boldsymbol{\theta}})$.

We apply this method to data on the Minimal Residual Disease (MRD) measured longitudinally during the treatment of childhood leukemia. Data are available on 240 cases treated according to 3 different protocols in the framework of an international clinical trial described elsewhere [4]. The MRD is an interval-censored measure of the residual presence of leukemic cells in the bone marrow. It was measured at 9 time-points to investigate its relationship with the disease process. Focus is on the probability of relapse, being the relapse an MRD measurement above a relevant threshold.

3. Results

We fitted models with different covariance structure. Table 1 summarises the results and shows the computed values of the Bayesian Information Criterion (BIC):

Table 1: Log-maximized likelihood and corresponding BIC for various models.

	Model	Covariance structure $\mathbf{V}_i$	$l(\hat{\boldsymbol{\theta}}; \mathbf{y})$	$\dim(\boldsymbol{\theta})$	BIC
(I)	mean only	$\sigma_\varepsilon^2 \mathbf{I}$	-1261.329	4	2225.030
(II)	(I) + random intercept	$\sigma_1^2 \mathbf{J} + \sigma_\varepsilon^2 \mathbf{I}$	-1140.217	5	2317.149
(III)	(I) + random slope	$\sigma_2^2 \mathbf{t} \mathbf{t}^T + \sigma_\varepsilon^2 \mathbf{I}$	-1127.968	5	2292.651
(IV)	(I) + random effects uncorrelated	$\mathbf{Z} \mathbf{B} \mathbf{Z}^T + \sigma_\varepsilon^2 \mathbf{I}$, $\mathbf{B}$ diagonal	-1098.553	6	2241.164
(V)	(I) + random effects correlated	$\mathbf{Z} \mathbf{B} \mathbf{Z}^T + \sigma_\varepsilon^2 \mathbf{I}$	-1092.042	7	2235.485
(VI)	(I) + serial correlation	$\boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I}$	-1140.275	6	2324.608
(VII)	(II) + serial correlation	$\sigma_1^2 \mathbf{J} + \boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I}$	-1140.162	7	2331.725
(VIII)	(III) + serial correlation	$\sigma_2^2 \mathbf{t} \mathbf{t}^T + \boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I}$	-1077.028	7	2205.457
(IX)	complete model	$\mathbf{Z} \mathbf{B} \mathbf{Z}^T + \boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I}$	-1076.273	9	2218.633

The best fit is that of model VIII, which allows for the presence of both a random slope and a Gaussian process (the same results would be obtained using the AIC). We applied this model to calculate the probability of relapse. Table 2 reports the marginal probability of relapse estimated at relevant time-points.

Table 2: Marginal probability (%) of observing an MRD value on or above the relapse threshold at relevant times.

Treatment group	days from diagnosis						
	35	84	154	196	365	560	728
Standard Risk	3.2	3.6	4.8	5.9	12	19	24
Intermediate Risk	3.2	3.6	4.8	5.9	12	19	31
High Risk	34	35	36	37	40	42	44

The model allows also the computation of the conditional probability of observing an MRD value on or above the threshold, having previously observed a certain MRD history. For example, the probability of relapse at 365 days, for a patient treated according to the High Risk protocol is 44.7%, 95% C.I. (41.7, 47.4) if the previous MRD values are stable on high (bad) values. Corresponding figures if MRD history shows a slow or rapid improvement to low values, are 5.2%, 95% C.I. (2.8, 8.3) and 2.2%, 95% C.I. (0.7, 4.7), respectively.

4. Conclusions

This work aimed at adapting the general model for the analysis of longitudinal data to a very peculiar type of repeated measurements, characterized by interval censoring and collected within the framework of a survival study. Application to a real dataset shows that the approach is promising, as it allows a proper description of the data while maintaining flexibility of the method, feasibility of the computations and interpretability of the results. The profile in time of the relapse probability reflects the impact of the updated biomarker value on the disease course.

References

- [1] Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition, J. Wiley & Sons, New York.
- [2] Jacqmin-Gadda, H., Thiébaud, R., Chêne, G. and Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, **1**, 355-368.
- [3] Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd edition, Oxford University Press, Oxford.
- [4] Van Dongen, J.J.M. Seriu, T. Panzer-Gruemayer, E.R. Biondi A. *et al.* (1998). Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. *The Lancet*, **352**, 1731-1738.