A JOINT MODEL FOR LONGITUDINAL AND SURVIVAL DATA BASED ON AN $\mathrm{AR}(1)$ LATENT PROCESS

Silvia Bacci[1], Francesco Bartolucci[1] and Silvia Pandolfi[2]

[1] Department of Economics, University of Perugia, Perugia, Italy, (e-mail: silvia.bacci@unipg.it, francesco.bartolucci@unipg.it)

[2] Department of Political Sciences, University of Perugia, Perugia, Italy, (e-mail: pandolfi@stat.unipg.it)

KEYWORDS: EM algorithm; generalized linear models; non-ignorable missing; shared-parameter models.

1 Introduction

A relevant problem in the analysis of longitudinal data is due to missing observations, in particular when the missing mechanism is non-ignorable (Little & Rubin, 2002). In the statistical literature there exist different approaches to model such a mechanism. Here we focus on the shared-parameter approach (Wu & Carroll, 1988), which introduces random effects to capture the association between the measurement and the missing process. The idea is that there exists an underlying latent process, described by the random effects, that drives both observed processes. An example of shared-parameter approach is represented by Joint Models (JMs; Wulfsohn & Tsiatis, 1997; Henderson *et al.*, 2000; Tsiatis & Davidian, 2004; Rizopolous, 2012).

In the standard formulation, a JM is characterized by a generalized linear mixed model for the longitudinal process, with normally distributed random effects, and by a proportional hazard Cox's model (Cox, 1972) for the survival process, where the risk of the event of interest (e.g., death) at a given time depends on the expected value of the longitudinal response at the same time.

The standard JM formulation assumes the subject-specific random effects to be time constant. In order to relax this assumption, Bartolucci and Farcomeni (2014) introduce a family of mixed latent Markov models, where the non-ignorable missing process is accounted for through a discrete time-to-event history approach. Differently, Barrett *et al.* (2015) illustrate an approach for continuous longitudinal responses based on the discretization of time-to-event and on a hazard model formulated in terms of a probit model. In this way, exact likelihood inference is admitted for a wide range of random effects specifications.

In our contribution (Section 2), we propose to adopt a first-order autoregressive process, AR(1), instead of a discrete one, so that the resulting model is more parsimonious than that of Bartolucci and Farcomeni (2014) and we generalize the approach of Barrett *et al.* (2015) to different longitudinal outcomes, such as binary and count responses, using the quadrature method illustrated in Bartolucci *et al.* (2014) for parameter estimation. Our proposal is suitable for applications to data involving a range of different types of response (Section 3).

2 The proposed model

Aim of the work is to relax the hypotheses of the model of Barrett *et al.* (2015) by assuming a generalized linear parametrization for the longitudinal process and a sequence of random effects that follows an AR(1) process. The key-point is that each observation j for subject i is taken at time t_{ij} falling in a certain "time window" or period $s_{ij} = s(t_{ij})$.

The sub-model for the longitudinal process is formulated as

$$g(\mu_{ij}) = \alpha_{is_{ij}} + \mathbf{x}_{ij}^T \boldsymbol{\beta}, \tag{1}$$

with $g(\cdot)$ denoting a suitable link function, μ_{ij} denoting the conditional expected value of the outcome $y_{ij} = y_i(t_{ij})$, and $\mathbf{x}_{ij} = \mathbf{x}_i(t_{ij})$ being the corresponding vector of covariates that may include t_{ij} itself. The random intercept α_{is} depends on the time as follows:

$$\alpha_{is} = \alpha_{i,s-1}\rho + \eta_{is}\sqrt{1-\rho^2}, \quad s > 1,$$

with $\alpha_{i1} = \eta_{i1}$ and where $\rho = \text{cor}(\alpha_{is}, \alpha_{i,s-1})$. Moreover, the error terms η_{is} are independent and distributed as $N(0,\, \sigma^2)$.

The sub-model for the survival process is defined as follows:

$$\log \frac{p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})}{1 - p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})} = \alpha_{is}\gamma + \mathbf{w}_{is}^T \boldsymbol{\delta}, \tag{2}$$

with S_i corresponding to the number of periods that subject i survives and $\mathbf{w}_{is}$ denoting the vector of covariates that are operative at time s on the survival process. In practice, the model based on assumptions (1) and (2) generalize the proposal of Barrett *et al.* (2015) to a generic (i.e., continuous, binary, count) longitudinal outcome.

In order to compute the likelihood function of any model in the proposed class, we rely on a quadrature method based on an equally spaced grid of points and on a recursion developed in the hidden Markov literature (see Baum *et al.*, 1970). This likelihood function is characterized by individual components $p(\mathbf{y}_i, s_i, d_i|\mathbf{X}_i, \mathbf{W}_i)$ based on suitably marginalizing out $\boldsymbol{\alpha}_i$ from the following expression, where $\boldsymbol{\alpha}_i$ is the vector of random effects:

$$p(\mathbf{y}_i, s_i, d_i|\boldsymbol{\alpha}_i, \mathbf{X}_i, \mathbf{W}_i) = \left[\prod_{s=1}^{s_i-1} p(S_i > s|S_i \geq s, \alpha_{is}, \mathbf{w}_{is})\right]$$
$$\times p(S_i > s_i|S_i \geq s_i, \alpha_{is}, \mathbf{w}_{is})^{d_i} p(S_i = s_i|S_i \geq s_i, \alpha_{is}, \mathbf{w}_{is})^{1-d_i}$$
$$\times \prod_{j=1}^{j_i} p(y_{ij}|\alpha_{is_{ij}}, \mathbf{x}_{ij}).$$

In this expression, d_i is the final status of subject i, equal to 1 if subject i is alive at the end of the last period of observation and to 0 otherwise, s_i is the number of periods of observation, $\mathbf{y}_i$ is the observed vector of responses with j_i elements, and $\mathbf{X}_i$ and $\mathbf{W}_i$ are matrices of covariates with columns $\mathbf{x}_{ij}$ and $\mathbf{w}_{is}$, respectively.

Note that the method of Barrett *et al.* (2015), based on exact likelihood inference, is no longer applicable in our extended approach. By our method we also obtain the corresponding score vector with respect to the free parameters. Maximization of the likelihood function is based on a quasi-Newton algorithm in which the observed information matrix is obtained by a numerical method based on the score vector. On the basis of this matrix, we also obtain standard errors for the parameter estimates.

3 Application

We propose three applications of JM specified by equations (1) and (2) on certain datasets, which are characterized by different types of response variable:

- data concerning repeated measurements of lung function in cystic fibrosis patients, in which the continuous response variable corresponds to the percent forced expiratory volume;
- data concerning the count of yearly new skin cancers so as to analyze the effect of β-carotene for the prevention of non-melanoma skin cancer;
- data concerning the follow up of 312 randomized patients with primary biliary cirrhosis, in which the longitudinal outcomes are given by the serum bilirubin levels (mg/dL) and by the presence of edema.

Acknowledgments

Authors acknowledge the financial support from the grant FIRB ("Futuro in ricerca") 2012 on "Mixture and latent variable models for causal inference and analysis of socio-economic data”, which is funded by the Italian Government (RBFR12SHVV).

References

BARRETT, J., DIGGLE, P., HENDERSON, R., & TAYLOR-ROBINSON, D. 2015. Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society, Series B.*, **77**, 131–148.

BARTOLUCCI, F., & FARCOMENI, A. 2014. A discrete time event-history approach to informative drop-out in multivariate latent Markov models with covariates. *Biometrics*, **71**, 80-89.

BARTOLUCCI, F., BACCI, S., & PENNONI, F. 2014. Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 267–288.

BAUM, L. E., PETRIE, T., SOULES, G., & WEISS, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**, 164–171.

COX, D. 1972. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B.*, **34**, 187–220.

FOLLMANN, D., & WU, M. 1995. An approximate generalized linear model with random effects for informative missing data. *Biometrics.*, **51**, 151–168.

HENDERSON, R., DIGGLE, P., & DOBSON, A. 2000. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

LITTLE, R. J. A., & RUBIN, D. B. 2002. *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

RIZOPOLOUS, D. 2012. *Joint models for longitudinal and time-to-event data with applications in R.* Boca Raton, FL: Chapman&Hall/CRC Press.

TSIATIS, A. A., & DAVIDIAN, M. 2004. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809–834.

WU, M., & CARROLL, R. 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.

WULFSOHN, M. S., & TSIATIS, A. A. 1997. A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.