bild: a package for BInary Longitudinal Data

M.Helena Gonçalves[1,2,⋆], **M.Salomé Cabral**[1,3], **Adelchi Azzalini**[4]

1. Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
2. Departamento de Matemática, FCT, Universidade do Algarve, Portugal
3. Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Portugal
4. Dipartimento di Scienze Statistiche, Università di Padova, Italy

⋆Contact author: mhgoncal@ualg.pt

Keywords: Binary longitudinal data, Exact likelihood, Marginal models, Markov Chain, Random effects.

The software tools that we propose are aimed at the analysis of binary longitudinal data from the point of view of likelihood inference, which requires complete specification of a stochastic model for the individual profile. Denote by y_{it} $(t = 1, \ldots, T_i) \in \{0, 1\}$ the response value at time t from subject i $(i = 1, \ldots, n)$, and by Y_{it} its generating random variable whose mean values is $\mathbb{P}\{Y_{it} = 1\} = \theta_{it}$. Associated to each observation time and each subject, a set of p covariates, x_{it}, is available. In our formulation the parameter of interest is the marginal probability of success, that is related to the covariates via a logistic regression model,

$$\operatorname{logit} \mathbb{P}\{Y_{it} = 1\} = x_{it}^{\top}\beta. \tag{1}$$

The dependence structure of the process corresponds to a second order Markov Chain. This set-up leads to consideration of the joint distribution of three components of the process at the time, (Y_{t-2}, Y_{t-1}, Y_t) say. Our choice is to impose the constraints

$$OR(Y_{t-1}, Y_{t-2}) = \quad \psi_1 \quad = OR(Y_{t-1}, Y_t) \tag{2}$$

$$OR(Y_{t-2}, Y_t | Y_{t-1} = 0) = \quad \psi_2 \quad = OR(Y_{t-2}, Y_t | Y_{t-1} = 1). \tag{3}$$

where ψ_1 and ψ_2 are two positive parameters. In the context of binary processes, dependence is more conveniently measured by odds ratios rather than correlations, and conditions (2)–(3) provide a parametrisation whose interpretation is similar to the partial autocorrelation of a Gaussian process, transferred to the odd-ratio scale. The problem is finding the $p_{hj} = \mathbb{P}\{Y_t = 1 | Y_{t-2} = h, Y_{t-1} = j\}, h, j = 0, 1$, satisfying the above-stated conditions, see [1] for details. This software allows the presence of individual random effects by adding the component $b_i \sim N(0, \sigma^2)$ in (1) leading to the logistic model with random intercept

$$\operatorname{logit} \mathbb{P}\{Y_{it} = 1\} = x_{it}^{\top}\beta + b_i, \tag{4}$$

where the b_i's are assumed to be sampled independently from each other. We reparameterise $\omega = \log \sigma^2$ both for numerical convenience and to improve accuracy of the asymptotic approximation to the distribution of MLEs. One dimensional integrals are computed using adaptive Gaussian quadrature. A frequent problem with longitudinal studies is the presence of missing data, this software allows for a simple pattern of missing data, details available in [2]. We considered a form of residuals of a fitted model, to be used for diagnostic purposes. Graphical analysis of residuals is difficult even for the simple case of logistic regression of binary data, due to the extreme discreteness of the binary data. To alleviate the problem of discreteness, we aggregate residuals across individuals at each given time point. The package, called **bild**, is a S4-methods package and provides R functions for parametric and graphical analysis of binary longitudinal data. The functions of **bild** have been written in R language, except for some FORTRAN routines which are interfaced through R. The main function performs the fit of parametric models via likelihood methods. Serial dependence and random effects are allowed according to the stochastic model chosen: independence, MC1 (1st order Markov Chain), MC2 (2nd order Markov Chain), MC1R (1st order Markov Chain with random effects) or MC2R (2nd order Markov Chain with random effects). Missing values and unbalanced data are automatically accounted for computing the likelihood function. Six plots are available in plot methods: Residuals vs Fitted, Residuals vs Time, ACF residuals, PACF residuals, Parametric fit and Individual mean profiles.

References

[1] M. Helena Gonçalves and Adelchi Azzalini (2008). Using Markov chains for marginal modelling of binary longitudinal data in an exact likelihood approach. *Metron*, LXVI, 157–181.

[2] M. Helena Gonçalves (2002). *Likelihood methods for discrete longitudinal data.* PhD thesis, University of Lisbon.