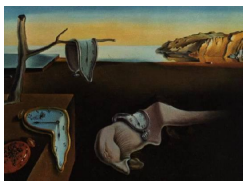

Journal of Statistical Software
February 2012, Volume 46, Issue 9. http://www.jstatsoft.org/

The R Package bild for the Analysis of Binary Longitudinal Data

M. Helena Gonçalves
Universidade do Algarve

M. Salomé Cabral
Universidade de Lisboa

Adelchi Azzalini
Università di Padova

Abstract

We present the R package **bild** for the parametric and graphical analysis of binary longitudinal data. The package performs logistic regression for binary longitudinal data, allowing for serial dependence among observations from a given individual and a random intercept term. Estimation is via maximization of the exact likelihood of a suitably defined model. Missing values and unbalanced data are allowed, with some restrictions. The code of **bild** is written partly in R language, partly in Fortran 77, interfaced through R. The package is built following the S4 formulation of R methods.

Keywords: binary longitudinal data, exact likelihood, marginal models, Markov chain, odds ratio, random effects.

1. Introduction

This paper describes the R (R Development Core Team 2011) package **bild** (Gonçalves, Cabral, and Azzalini 2012) available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=bild` for the parametric and graphical analysis of binary longitudinal data. Similarly to the generalized estimating equations (GEE) approach when this is applied to binary response data, the present methodology works by introducing a parametric model for the marginal distribution of the response variable but it differs from the GEE approach on another front, in that the parametric analysis developed here is associated to a fully specified stochastic model for the individual profiles.

Important work on the likelihood approach for discrete longitudinal data, with emphasis on the important special case of binary response, has been done by Fitzmaurice and Laird (1993), Fitzmaurice, Laird, and Rotnitzky (1993), and Fitzmaurice, Laird, and Lipsitz (1994), via the so-called mixed-parametrization. However their approach is unsuitable to handle series of different length of response across individuals, and the interpretation of the association parameters is somewhat problematic; see the discussions of these papers.

An alternative likelihood-based formulation for a logistic regression model which allows for the presence of serial dependence has been presented by Azzalini (1994); its adaptation to the case of longitudinal data is immediate. Similarly to some transitions models, this formulation assumes that serial dependence is regulated by a Markov chain mechanism, but the distinctive fact here is that a suitable parametrization is adopted so that the logistic regression parameters preserve their meaning as if we were working with in the traditional case of independent observations. In other words, modeling of the mean value and of serial dependence are effectively separated. The proposal has been implemented in the S-PLUS package described by Azzalini and Chiogna (1997).

The aforementioned approach has been developed by Gonçalves (2002) and Gonçalves and Azzalini (2008), in various directions: (i) extension from first order to second order Markov dependence, (ii) allowance of missing data, (iii) inclusion of a random intercept term in the linear predictor. Section 2 provides the minimal theoretical background to follow Section 3 which introduces to the practical use of the software.

The functions of **bild** have been written in R language (R Development Core Team 2011), with some Fortran 77 routines which are interfaced through R. The package is built following the S4 formulation of R methods.

2. Parametric models for binary data

2.1. Binary Markov chains

Denote by $y_{it} \in \{0, 1\}$ the binary response value at time t ($t = 1, \ldots, T_i$) from subject i ($i = 1, \ldots, n$), and by Y_{it} its generating random variable whose mean value is $\mathsf{P}(Y_{it} = 1) = \theta_{it}$. For each observation time and for 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,

$$\operatorname{logit} \theta_{it} = x_{it}^\top \beta \tag{1}$$

where β is a p-dimensional parameter. The dependence structure of the process is taken to be a second order Markov chain, which is suitably parameterized so that the marginal parameter β retains its meaning irrespective of the serial dependence.

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). To simplify notation, we drop the subscript i where it is not essential. Our choice is to impose the constraints

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

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

where the notation $OR(U, V)$ denotes the odds ratio of the joint distribution of a pair of binary random variables (U, V), and ψ_1 and ψ_2 denote two positive parameters. In the context of binary processes, dependence is more conveniently measured by odds ratios than by correlations, and conditions (2)–(3) provide a parametrization whose interpretation is similar to the partial autocorrelation of a Gaussian process, transferred to the odds ratio scale. The algebraic problem is finding the transition probabilities

$$p_{hj} = \mathsf{P}(Y_t = 1 | Y_{t-2} = h, Y_{t-1} = j), \qquad h, j = 0, 1, \tag{4}$$

satisfying the above-stated conditions; see Gonçalves and Azzalini (2008) for more details and Gonçalves (2002) for a full account.

Therefore in the adopted formulation, interest lies in the marginal probability of success, and this is related to the covariates via the logistic regression model (1); hence β is the parameter of interest. Serial dependence is regulated by $\lambda = (\lambda_1, \lambda_2) = (\log \psi_1, \log \psi_2)$, which we assume to be constant across time and individuals. Notice that, when $\lambda_2 = 0$, the Markov chain reduces effectively to first order dependence, and we recover the formulation of Azzalini (1994).

2.2. Likelihood inference

We shall now consider likelihood inference based on the a sample of n individual profiles, assumed to be independent from each other. The contribution from a generic individual to the log-likelihood for the parameters (β, λ) is

$$\begin{aligned} \ell^F(\beta, \lambda) \quad = \quad & \big[y_1 \operatorname{logit}(\theta_1) + \log(1-\theta_1)\big] + \big[y_2 \operatorname{logit}(p'_{y_1}) + \log(1-p'_{y_1})\big] \; + \\ & \sum_{t=3}^{T} \Big[y_t \operatorname{logit}(p_{y_{t-2},y_{t-1}}) + \log(1-p_{y_{t-2},y_{t-1}}) \Big] \end{aligned} \tag{5}$$

where the three blocks on the right-hand side represent the contribution to the log-likelihood from y_1, y_2, and $(y_3, \ldots, y_T)$, respectively, and where $p'_j = \mathsf{P}(Y_t = 1 | Y_{t-1} = j)$. The overall log-likelihood function is obtained as the sum of the n individual contributions of type (5).

Maximization of the log-likelihood must be performed by numerical methods. To improve speed of convergence, expressions of the score functions have been obtained; see the references quoted in Section 2.1.

2.3. Residuals

We introduce a form of residuals of a fitted model, to be used for diagnostic purposes. Denote by $Y^{(s)}$ the set of random variables $\{Y_1, \ldots, Y_s\}$ representing the portion of the i-th individual profile up to time s. The conditional mean given the past observations is denoted by

$$m_t = \mathsf{E}(Y_t | Y^{(t-1)})$$

for $t = 2, 3, \ldots$, and $t = 1$ does not involves any conditioning, so $m_1 = \mathsf{E}(Y_1)$. Then define the standardized residual for $t = 1, 2, \ldots$

$$r_t = \frac{y_t - m_t}{\sqrt{m_t\,(1-m_t)}}.$$

Graphical analysis of residuals is difficult even in the simple case of logistic regression for independent data, due to the extreme discreteness of binary data. To alleviate the problem of discreteness, and at the same time to reduce the number of plots to examine, it is sensible to aggregate the residuals across individuals, at each given time point, in the form

$$R_t = \frac{\sum_i (y_{it} - m_{it})}{\{\sum_i m_{it}(1-m_{it})\}^{1/2}} \tag{6}$$

where the index i, now re-introduced, runs across the whole set of n individuals, or possibly an homogeneous sub-group of them.

2.4. Missing data

A frequent problem with longitudinal studies is the presence of missing data, since it is difficult to have complete records of all individuals, especially in cases when measurements are taken at occasions very distant in time. Missing values are allowed on the response, provided they are missing at random in the terminology of Little and Rubin (1987).

If missing data occur at the beginning or at the end of an individual profile, this poses no problems, since this case is equivalent to a designed unbalance in the length profile T_i for that individual. Some restrictions exist for the presence of missing data when they occur in the middle of the profile. If the first order dependence model is considered the present implementation requires that only one missing value should appear between two significant observations. For the second order dependence model we need that the pattern of missing observations satisfies the requirement that missing data have two observed values on each side of the time sequence, except for the two end portions of the observation period, where no restriction is made. Therefore, if there is a missing value at time point $t-2$, it is required that there are observations at time points $t-4, t-3, t-1, t$.

In practice, the program performs the fit even when the above conditions are not fulfilled. In these cases, however, a warning message is printed since the log-likelihood is not computed exactly, with consequent slight inaccuracy of the results.

2.5. Random effects

Individual random effects $b_i \sim N(0, \sigma^2)$ can be incorporated as an additive term to the linear predictor in (1), leading to the logistic model with random intercept

$$\operatorname{logit} \mathsf{P}(Y_{it} = 1|b_i) = x_{it}^\top \beta + b_i, \qquad (i = 1, \ldots, n) \tag{7}$$

where the b_i's are assumed to be sampled independently from each other.

The corresponding expression for the contribution of the i-th subject to the likelihood function is obtained by integrating over the distribution of b_i in the expression of the likelihood for the fixed effect model evaluated at $\beta^{(b_i)}$, which is the same of β with the intercept β_0 replaced by $\beta_0 + b_i$. It is convenient to reparametrize $\omega = \log \sigma^2$ both for numerical convenience and to improve accuracy of the asymptotic approximation to the distribution of maximum likelihood estimates (MLE's). In explicit terms, the likelihood for the random intercept case is

$$L_i^R(\beta, \lambda, \omega) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{\mathbb{R}} L_i^F(\beta^{(b_i)}, \lambda|b_i) \exp\left(-\frac{b_i^2}{2\,\sigma^2}\right) \mathrm{d}b_i \tag{8}$$

where

$$L_i^F(\beta^{(b_i)}, \lambda|b_i) = \exp\{\ell_i^F(\beta^{(b_i)}, \lambda)\}$$

is computed from (5). Clearly, the log-likelihood for the whole sample is given by

$$\ell^R(\beta, \lambda, \omega) = \sum_{i=1}^{n} \log L_i^R(\beta, \lambda, \omega).$$

In practice the integrals in (8) are computed using adaptive Gaussian quadrature. To improve efficiency of the numerical optimization of the log-likelihood, it is convenient to make use of its derivatives. See Gonçalves and Azzalini (2008) for more details and Gonçalves (2002) for a full description.

In most generalized linear mixed model (GLMM) formulations for binary data, an expression of type (7) is also used, but assuming independence of observations within a given individual, conditionally on b_i. Since here we allow for the presence of serial dependence, the methodology can be used to test the assumption of conditional independence underlying a GLMM model, by examining the estimates of λ_k's and the associated quantities.

In interpreting the parameters β of (7), one must bear in mind that the inclusion of the random effect b_i, as given by (7), alters the meaning of the β's with respect to their meaning in a model with fixed effects only, since clearly

$$\mathsf{E}\left(\frac{e^{\eta+b}}{1+e^{\eta+b}}\right) \neq \frac{e^{\eta}}{1+e^{\eta}} \tag{9}$$

where $\eta = x_{it}^{\top}\beta$ denotes the systematic part of the linear predictor.

To obtain the exact value on the left-hand side an integration is required, in practice via numerical methods. To avoid this integration, a simple and effective approximation has been considered by Zeger, Liang, and Albert (1988), which in our case amounts to evaluating the right-hand side of (9) with η replaced by $a(\omega)\,\eta$, where

$$a(\omega) = (c^2 e^{\omega} + 1)^{-1/2} \tag{10}$$

and $c = 16\sqrt{3}/(15\pi)$. This adjustment is incorporated by the package in the graphical display of estimated probabilities, and referred to as "adjusted fit" in Section 3.1.

In addition to estimation of β and λ, it is of interest to obtain estimates $\hat{b}_i$ of the individual random effects, b_i's. The appropriate quantity to consider is the conditional expectation of b_i given the observed value of the i-th individual profile y_i, that is $\mathsf{E}(b_i|y_i;\beta,\lambda)$, but its exact computation is difficult. A simple alternative follows naturally from the following argument: if the parameters β of the systematic component $\eta_i = x_{it}^{\top}\beta$ of (7) were available, one could estimate b_i by fitting a simple logistic model, separately for each individual, regarding η_i as a fixed constant. In practice one replaces β by its estimate to compute η_i, and then fits a logistic regression model to y_i, with η_i treated as an "offset".

Once estimates $\hat{b}_i$ of the individual random effects are available, one can use the estimated conditional linear predictor $\hat{b}_i + x_{it}^{\top}\hat{\beta}$ to compute transition probabilities. This process will deliver an approximation to the conditional mean

$$m_{it} = \mathsf{E}(Y_{it}|Y_i^{(t-1)}, b_i) \tag{11}$$

which in turns allows computation of residuals r_{it}, along the lines of Section 2.3. See Gonçalves and Azzalini (2008) for details.

3. Using package bild

3.1. Package overview

The package is built around its main function `bild()` which performs the fit of models of type described in the previous section by maximizing the log-likelihood according to some serial dependence structure (5).

Serial dependence of first order and second order Markovian type will be identified as `MC1` and `MC2`, respectively; the corresponding parameters $\lambda_1 = \log \psi_1$ and $\lambda_2 = \log \psi_2$ are denoted `log.psi1` and `log.psi2`. When a random intercept term is included, then notation for the dependence structure will be either `MC1R` and `MC2R`, depending of the order of serial dependence. In addition the dependence `ind` is allowed, to select independence.

The arguments used in a call to the function `bild()` are:

```
bild(formula, data, time, id, subSET, aggregate, start, trace,
  dependence = "ind", method = "BFGS", control = bildControl(),
  integrate = bildIntegrate())
```

We summarize next the main arguments of `bild()` plus two auxiliary functions, `bildControl()` and `bildIntegrate()`, which help to regulate the working of the main function.

`formula`: Description of the model to be fitted of the form `response ~ predictors`.

`data`: `data.frame` whose structure is described in Section 3.2. Notice that certain components of `data` influence the working of the function.

`time`: String that matches the name of the `time` variable in `data`. By default, the program expects a variable named `time` to be present in the `data.frame`, otherwise the name of the variable playing the role of `time` must be declared by assigning `time` here.

`id`: String that matches the name of the `id` variable in `data`. By default, the program expects a variable named `id` to be present in the `data.frame`, otherwise the name of the variable playing the role of `id` must be declared by assigning `id` here.

`subSET`: Optional expression indicating the subset of `data` that should be used in the fit. This is a logical statement of the type `(variable1 == "a" & variable2 > x)` which identifies the observations to be selected. All observations are included by default.

`aggregate`: String that identifies the levels of the factor to perform the parametric fit in the `plot` method.

`start`: Starting values for the optimization can be defined through the argument `start`. The values in `start` vector depends on the structure of the dependence model and not on the parameters of model predictor. The structure of the vector `start` should be: (λ_1) when `dependence = "MC1"`, (λ_1, λ_2) when `dependence = "MC2"`, (λ_1, ω) when `dependence = "MC1R"` or $(\lambda_1, \lambda_2, \omega)$ when `dependence = "MC2R"`.

`trace`: By default set to `FALSE`. If `trace = TRUE`, the intermediate values of the likelihood parameters and of the likelihood itself are printed.

`dependence`: Expression stating which dependence structure should be used in the fit. The default value is `"ind"`. According to the stochastic model chosen serial dependence and random effects are allowed. There are five options: `"ind"` (independence), `"MC1"` (first order Markov chain), `"MC2"` (second order Markov chain), `"MC1R"` (first order Markov chain with random intercept) or `"MC2R"` (second order Markov chain with random intercept).

`method:` The method to be used in the optimization process: `"BFGS"`, `"CG"`, `"L-BFGS-B"` and `"SANN"`. The default is `"BFGS"`. See `optim` for details.

`control:` `bildControl()` returns a list of algorithmic constants for the optimizer `optim`, via a call of the form: `bildControl(maxit, abstol, reltol)`.

`integrate:` `bildIntegrate()` returns a list of constants that are used to compute integrals based on a Fortran 77 subroutine package `QUADPACK` for the numerical computation of definite one-dimensional integrals. The list is generated by a call of the form: `bildIntegrate(li, ls, epsabs, epsrel, limit, key, lig, lsg)`. For given values of `li` and `ls`, the above-described numerical integration is performed over the interval (`li`$\cdot\sigma$, `ls`$\cdot\sigma$) to compute the integral given by (8) where $\sigma = \exp(\omega/2)$ is associated to the current parameter value ω examined by the `optim` function. In some cases, this integration may generate an error, and the user must suitably adjust the values of `li` and `ls`. In case different choices of these quantities all lead to a successful run, it is recommended to retain the one with largest value of the log-likelihood. Integration of the gradient of (8) is regulated similarly by `lig` and `lsg`.

Six plots (selectable by `which`) are available in the `plot` method. By default, the first five are provided. The options are:

`which = 1` provides the plot of residuals vs. fitted values.

`which = 2` provides the plot of residuals vs. time.

`which = 3` provides the plot of ACF residuals.

`which = 4` provides the plot of PACF residuals.

`which = 5` provides the parametric fitted model if the dependence structure is `"ind"`, `"MC1"` or `"MC2"`. If the dependence structure is `"MC1R"` or `"MC2R"` the parametric adjusted fit is provided and the user can set `add.unadjusted = TRUE` to obtain the unadjusted fit.

`which = 6` provides individual mean profiles and is used only if the random intercept is present.

The `show-method` displays simple summary of a `bild()` object and the `summary-method` returns a more detailed list of summary statistics of the fitted model. Moreover, `bild-class` allows the user to extract several items produced by the maximum likelihood procedure. Besides the parameter estimates and other quantities featuring in the summary table of the fitted model, one can extract the residuals values, the fitted values and the transition probabilities. For a full description of the available quantities, see the list of slots of `bild-class` provided with the package documentation.

3.2. Data structure

The structure of the data is a `data.frame`. Each element of the `data` argument must be identifiable by a name. `NA` values can then be inserted in the response variable, provided that the missing values are missing at random (for details see Section 2.4). The response variable represent the individual profiles of each subject, it is expected a variable in the `data.frame`

that identifies the correspondence of each component of the response variable to the subject that it belongs, by default is named `id` variable. It is expected a variable named `time` to be present in the `data.frame`. The `time` variable should identify the time points that each individual profile has been observed.

If a response profile is replicated several times, a variable called `counts` must be created accordingly. This vector is used for weighting the response profile indicating for each individual profile the number of times that is replicated. The vector `counts` must repeat the number of the observed replications for each individual profile as many times as the number of observed time points for the correspondent profile. If each profile has been observed only once, the construction of the vector `counts` is not required.

In the two next subsections we illustrate the working of the package with the help of two real datasets.

3.3. Example: Locust data

The dataset has been presented and analyzed by MacDonald and Raubenheimer (1995), and subsequently examined by other authors, including Gonçalves and Azzalini (2008) using the methodology considered here. The data have been collected to study the effect of hunger on the locomotion behaviour of locusts (*Locusta migratoria*). Specifically 24 locusts have been observed at 161 time points, at thirty-second intervals; the subjects were divided in two treatment groups ("feed" and "unfeed"), and within each of the two groups, the subjects were alternatively "male" and "female". For the purpose of this analysis the categories of the response variable were "moving" and "not moving". To start analyzing the data we first load the package

```
R> library("bild")
```

The available data have the following structure:

```
R> str(locust)
```

```
'data.frame':        3864 obs. of  5 variables:
 $ id  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ move: num  0 0 0 0 0 0 0 0 0 0 ...
 $ sex : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ time: num  0.00833 0.01667 0.025 0.03333 0.04167 ...
 $ feed: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

where the `time` unit has been set equal to an hour, for numerical convenience. The function `bild()` is called to fit the model

$$\text{logit}(\theta_{it}) = \beta_0 + \beta_1\,\text{time} + \beta_2\,\text{time}^2 + \beta_3\,\text{feed} + \beta_4\,\text{time} \times \text{feed} + \beta_5\,\text{time}^2 \times \text{feed}$$

using a dependence structure `MC2` via the statements

```
R> locust2 <- bild(move ~ (time + I(time^2)) * feed, data = locust,
+    start = NULL, aggregate = feed, dependence = "MC2")
R> summary(locust2)
```

```
Call:
bild(formula = move ~ (time + I(time^2)) * feed, data = locust,
    aggregate = feed, start = NULL, dependence = "MC2")

Number of profiles in the dataset:  24
Number of profiles used in the fit:  24
Log likelihood:  -1562.763
AIC:  3141.527

Coefficients:
                Label      Value Std. Error t value  p-value
(Intercept)         1 -1.5302203 0.25469405  -6.008 0.000000
time                2  4.4317241 0.84767026   5.228 0.000000
I(time^2)           3 -2.4148937 0.59676345  -4.047 0.000052
feed1               4 -2.1531131 0.53309716  -4.039 0.000054
time:feed1          5 -4.8257679 1.63733357  -2.947 0.003205
I(time^2):feed1     6  4.2165054 1.09753049   3.842 0.000122
log.psi1            7  1.4588826 0.11192364  13.035 0.000000
log.psi2            8  0.9143597 0.09671381   9.454 0.000000

Message:  0
```

All the parameter estimates are significant at 5% level. This shows, among other things, that a quadratic time effect is present, and that a difference exists between the two groups ("feed" and "unfeed"). Moreover the estimates of $\lambda_1 = \log \psi_1$ and $\lambda_2 = \log \psi_2$ point strongly to second order serial dependence. To explore further this point, we can fit a similar model but adopting `MC1` dependence structure via

```
R> locust1 <- bild(move ~ (time + I(time^2)) * feed, data = locust,
+    start = NULL, aggregate = feed, dependence = "MC1")
```

and compute the likelihood ratio test to compare the two dependence structures

```
R>  2 * (getLogLik(locust2) - getLogLik(locust1))

[1] 88.45186
```

This change of deviance, compared with the χ^2_1 reference distribution, produces a p value about 0, confirming that the `MC2` structure is significantly preferable to `MC1`. The graphical display of the fitted probabilities as shown in Figure 1 can be obtained using the `plot` method (setting `which = 5`) by

```
R> plot(locust2, which = 5, ylab = "probability of locomotion")
```

The residual analysis can be summarized, as shown in Figure 2, using the `plot` method setting `which = 1` for residuals vs. fitted, `which = 2` for residuals vs. time, `which = 3` for ACF residuals and `which = 4` for PACF residuals via the following statements,

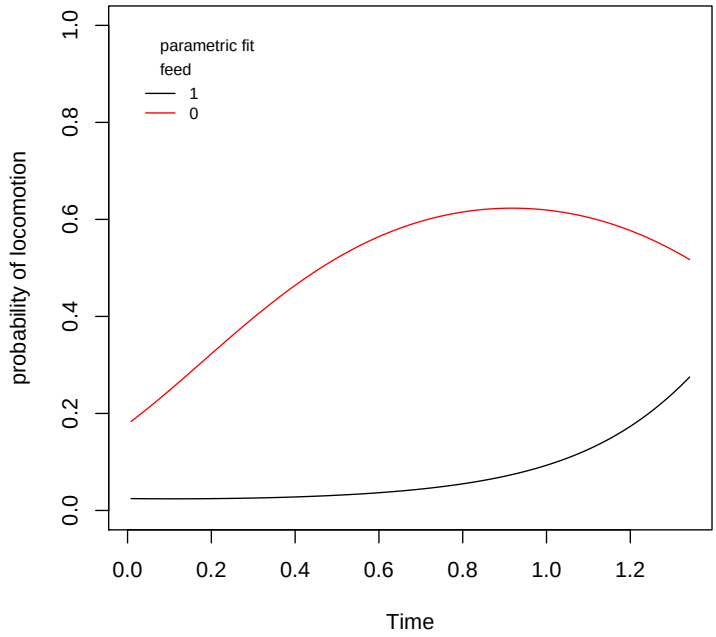

Figure 1: Probability of locomotion for the locust data under `MC2` assumption.

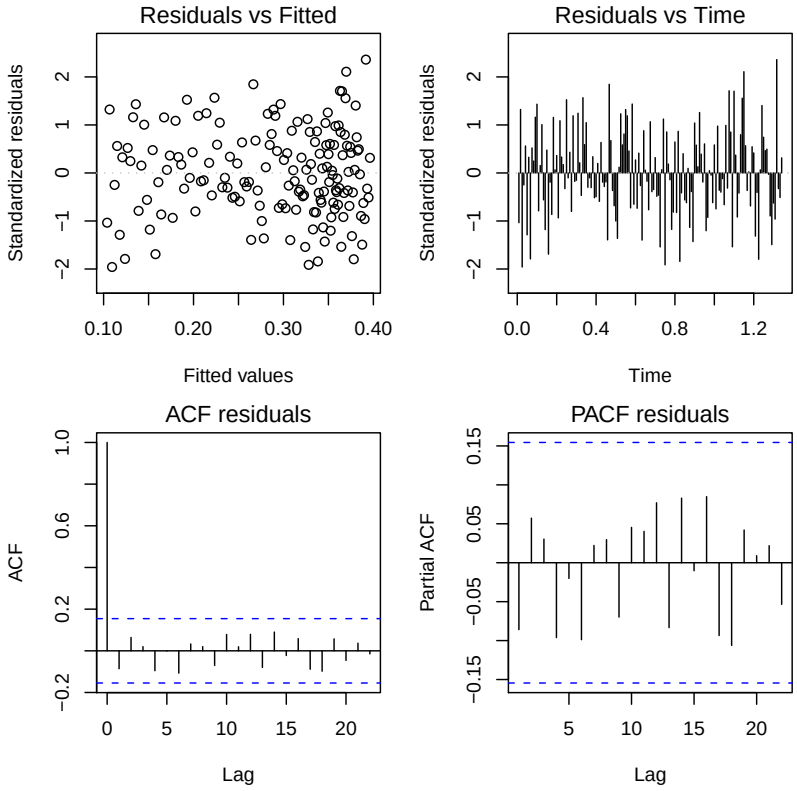

Figure 2: Residual plots of locust data under the `MC2` model.

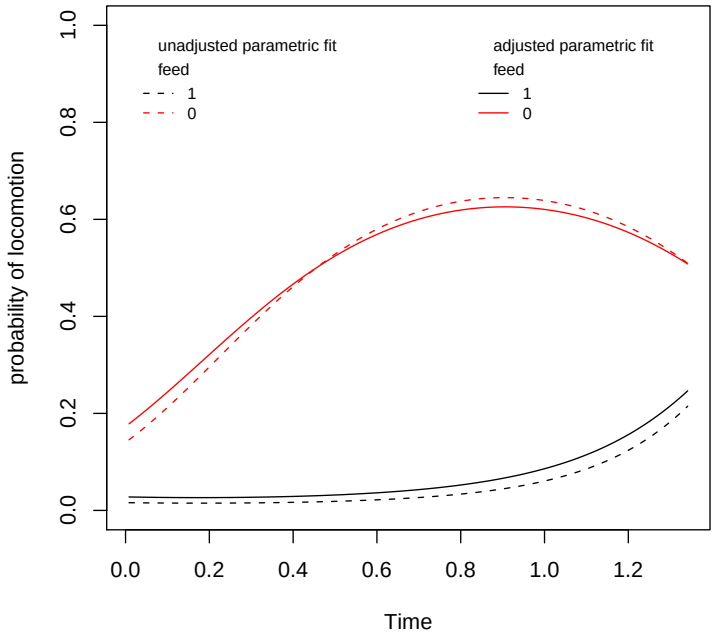

Figure 3: Probability of locomotion for the locust data under `MC2R` dependence.

```
R> plot(locust2, which = 1)
R> plot(locust2, which = 2)
R> plot(locust2, which = 3)
R> plot(locust2, which = 4)
```

For models with random intercept, `MC1R` and `MC2R`, the user may need to specify the `integrate` argument list changing the integration limits in the `bildIntegrate()` function.

If a random intercept is considered in the previous model (`dependence = "MC2"`) the call to the function `bild()` is now done setting `dependence = "MC2R"`. However, using the default settings for the integration limits of log-likelihood and its gradient, the call to `bild` generates an error. Therefore, the function `bildIntegrate()` was used to define new limits for likelihood and gradient as indicated here:

```
R> Integ <- bildIntegrate(li = -2.5, ls = 2.5, lig = -2.5, lsg = 2.5)
R> locust2r <- bild(move ~ (time + I(time^2)) * feed, data = locust,
+    start = NULL, aggregate = feed, dependence = "MC2R", integrate = Integ)
```

The fitted probabilities for the model given by `locust2r`, choosing `which = 5` in the `plot` method, can be adjusted (by default) or unadjusted (`add.unadjusted = TRUE`), see the Figure 3. The adjusted fits use the approximation considered by Zeger *et al.* (1988) and presented in equation (10). The unadjusted fits give the estimate probability of locomotion for the locust with the unobserved random intercept $b_{0i} = 0$, in both groups.

```
R> plot(locust2r, which = 5, ylab = "probability of locomotion",
+    add.unadjusted = TRUE)
```

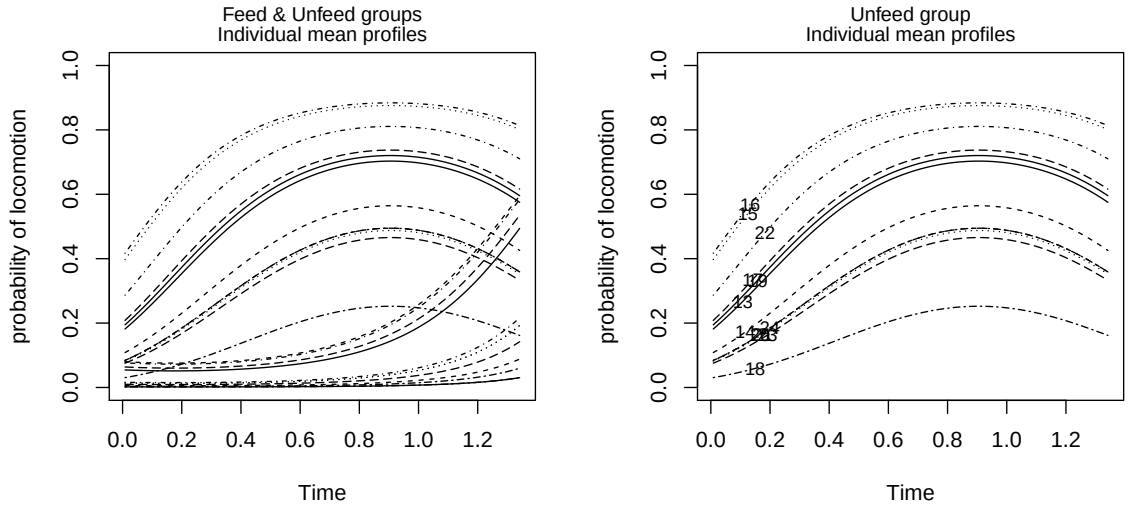

Figure 4: Individual mean profiles of locust data for `MC2R`.

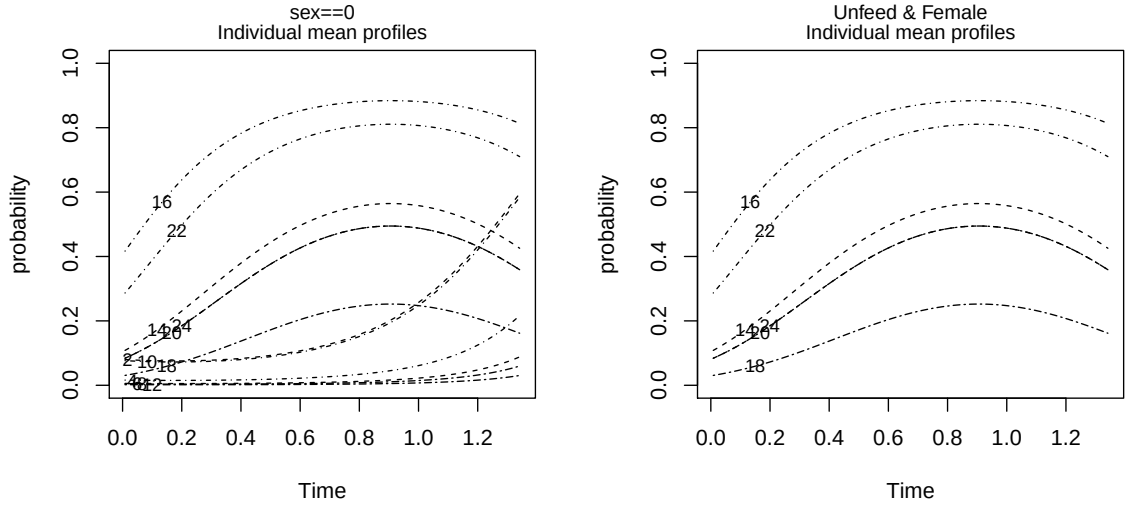

Figure 5: Individual mean profiles of locust data for `MC2R` using `subSET` option.

The individual mean profile for all subjects (by default) or to a subset (`subSET = ...`) can be obtained using the `plot` method choosing `which = 6`. The identification of the subjects is also allowed by setting `ident = TRUE`. These options are only available for random intercept models. See Figure 4 as the results of the settings below,

```
R> plot(locust2r, which = 6, ylab = "probability of locomotion",
+    main = "Feed & Unfeed groups")
R> plot(locust2r, which = 6, ident = TRUE, subSET = feed == "0",
+    ylab = "probability of locomotion", main = "Unfeed group")
```

In Figure 5 examples of individual mean profiles for females (`sex = 0`) are given for both treatments groups and for unfeed group (`feed = 0`) only. In these two plots the identification of the subjects is produced setting `ident = TRUE`.

```
R> plot(locust2r, which = 6, subSET = (sex == "0"), main = "sex==0",
+    ident = TRUE)
R> plot(locust2r, which = 6, subSET = (feed == "0" & sex == "0"),
+    main = "Unfeed & Female", ident = TRUE)
```

3.4. Example: Muscatine data

Fitzmaurice *et al.* (1994) and Azzalini (1994) have analyzed a subset of data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in school children from Muscatine (Iowa, USA). The data set contains records on 1014 children who were 7–9 years old in 1977 and were examined in 1977, 1979 and 1981. The binary response of interest is whether the child is obese (1) or not (0). A marginal model is appropriate to examine the probability of obesity as a function of gender and age. However, many data records are incomplete, since not all children participate in all the surveys.

The data structure is as follows:

```
R> str(muscatine)
```

```
'data.frame':        156 obs. of  5 variables:
 $ id    : int  1 1 1 2 2 2 3 3 3 4 ...
 $ obese : int  1 1 1 1 1 0 1 0 1 1 ...
 $ sex   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ time  : num  1 2 3 1 2 3 1 2 3 1 ...
 $ counts: num  20 20 20 7 7 7 9 9 9 8 ...
```

If we wish to decompose the `time` effect in orthogonal components, the following are the appropriate statements:

```
R> muscatine$time1 <- c(-1, 0, 1)
R> muscatine$time2 <- c(1, -2, 1)
```

The function `bild()` is called to fit the model:

$$\text{logit}(\theta_{it}) = \beta_0 + \beta_1\,\text{time1} + \beta_2\,\text{time2} + \beta_3\,\text{sex} + \beta_4\,\text{time1} \times \text{sex} + \beta_5\,\text{time2} \times \text{sex}$$

In order to investigate the effect of the presence of the random intercept in the model above `bild()` was called with `dependence = "MC2R"`. For this case we also make use of the `start` argument to assign initial values to the nuisance parameters, namely $(\lambda_1, \lambda_2, \omega) = (\log\psi_1, \log\psi_2, \log\sigma^2)$. Using the default settings for the integration limits of log-likelihood and its gradient, the call to `bild` generates NaN values for the standard errors. Again, the function `bildIntegrate()` was used to define new limits for gradient as indicated here:

```
R> Integ <- bildIntegrate(lig = -3.95, lsg = 3.95)
R> musc2r <- bild(obese ~ (time1 + time2) * sex, data = muscatine,
+    time = "time1", start = c(1, 1, 1), dependence = "MC2R",
+    integrate = Integ)
R> summary(musc2r)

Call:
bild(formula = obese ~ (time1 + time2) * sex, data = muscatine,
    time = "time1", start = c(1, 1, 1), dependence = "MC2R",
    integrate = Integ)

Number of profiles in the dataset:  52
Number of profiles used in the fit:  52
Log likelihood:  -947.2375
AIC:  1912.475

Coefficients:
            Label       Value Std. Error t value  p-value
(Intercept)     1 -3.12976020  0.9461909  -3.308 0.000940
time1           2  0.30762161  0.1771379   1.737 0.082454
time2           3  0.03864388  0.0774980   0.499 0.618030
sex1            4  0.07612533  0.3025062   0.252 0.801313
time1:sex1      5  0.36197649  0.2254992   1.605 0.108445
time2:sex1      6 -0.19804757  0.1219094  -1.625 0.104259
log.psi1        7  0.52345227  1.4620159   0.358 0.720317
log.psi2        8  0.06634889  1.4539330   0.046 0.963602

Random effect (omega):
     Value Std. Error
 2.4307357  0.7450833

Message:  0
```

The results of `summary(musc2r)` suggest that there is a linear increase (on the logit scale) in the rate of obesity over time, with no statistically discernible differences between males and females. Also, a simpler correlation structure seems to be appropriate. So, we consider the model with random intercept but with the first order dependence and only with the linear effect over time (`time1`).

```
R> musc1r <- bild(obese ~ time1, data = muscatine, time = "time1",
+    start = c(1, 1), dependence = "MC1R")
R> summary(musc1r)

Call:
bild(formula = obese ~ time1, data = muscatine, time = "time1",
    start = c(1, 1), dependence = "MC1R")

Number of profiles in the dataset:  52
```

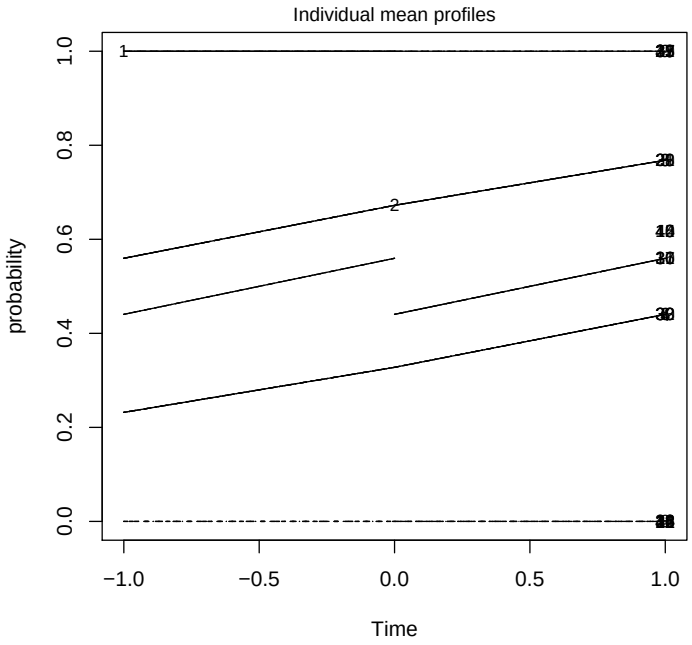

Figure 6: Individual mean profiles for Muscatine data with `MC1R` structure dependence.

```
Number of profiles used in the fit:  52
Log likelihood:  -950.9244
AIC:  1909.849

Coefficients:
            Label      Value Std. Error t value  p-value
(Intercept)     1 -3.0992732  0.3590851  -8.631 0.00000
time1           2  0.4790637  0.1123400   4.264 0.00002
log.psi1        3  0.4292217  0.5310888   0.808 0.41898

Random effect (omega):
     Value Std. Error
 2.4352198  0.2728933

Message:  0
```

The decrease in deviance between the models given by `musc1r` and `musc2r` is $\Delta D = 2 \times (950.92 - 947.24) = 7.37$ on five degrees of freedom (p value$= 0.19429$). Thus, the model with only the linear effect over time and `MC1R` is not rejected at the level of significance 5%, the results of `summary(musc1r)` confirms the linear increase (on the logit scale) in the rate of obesity over time. To obtain the individual means profile we set `which = 6` in the `plot` method and the result is obtained in Figure 6.

```
R> plot(musc1r, which = 6, ident = TRUE)
```

4. Closing remarks

In this paper we present an overview of the **bild** package for the analysis of binary longitudinal data. The theory used for model fitting is summarized briefly, and the functions of the package are described in detail. Practical use of **bild** is illustrated for the case of two real examples. A substantial computational burden is involved by the numerical integration connected to the random effects of Section 2.5, but this is not heavier than other formulations which incorporate random effects in discrete longitudinal data when a similar exact numerical integration is performed.

Acknowledgments

We would like to thank the referees and one of the Editors for a number of comments which have lead to a much improved presentation of the material. Research partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal-FCT under the project (PEst-OE/MAT/UI0006/2011).

References

Azzalini A (1994). "Logistic Regression for Autocorrelated Data with Application to Repeated Measures." *Biometrika*, **81**(4), 767–775. Amendment: Volume 84 (1997), 989.

Azzalini A, Chiogna M (1997). "S-PLUS Tools for the Analysis of Repeated Measurements Data." *Computational Statistics*, **12**, 53–66.

Fitzmaurice GM, Laird NM (1993). "A Likelihood-Based Method for Analyzing Longitudinal Binary Responses." *Biometrika*, **80**(1), 141–151.

Fitzmaurice GM, Laird NM, Lipsitz SR (1994). "Analyzing Incomplete Longitudinal Binary Responses: A Likelihood-Based Approach." *Biometrics*, **50**, 601–612.

Fitzmaurice GM, Laird NM, Rotnitzky AG (1993). "Regression Models for Discrete Longitudinal Responses." *Statistical Science*, **8**(3), 284–309.

Gonçalves MH (2002). *Likelihood Methods for Discrete Longitudinal Data*. Ph.D. thesis, University of Lisbon.

Gonçalves MH, Azzalini A (2008). "Using Markov Chains for Marginal Modelling of Binary Longitudinal Data in an Exact Likelihood Approach." *Metron*, **LXVI**, 157–181.

Gonçalves MH, Cabral MS, Azzalini A (2012). ***bild**: A Package for BInary Longitudinal Data*. R package version 1.0-4, URL http://CRAN.R-project.org/package=bild.

Little RJA, Rubin DB (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

MacDonald IL, Raubenheimer D (1995). "Hidden Markov Models and Animal Behaviour." *Biometrical Journal*, **37**, 701–712.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Zeger SL, Liang KY, Albert PS (1988). "Models for Longitudinal Data: A Generalized Estimating Equation Approach." *Biometrics*, **44**, 1049–1060.

Affiliation:

M. Helena Gonçalves
Centro de Estatística e Aplicações da Universidade de Lisboa
Departamento de Matemática, FCT, Universidade do Algarve
Gambelas, Portugal
E-mail: mhgoncal@ualg.pt

M. Salomé Cabral
Centro de Estatística e Aplicações da Universidade de Lisboa
Departamento de Estatística e Investigação Operacional, FCUL
Lisboa, Portugal
E-mail: salome@fc.ul.pt

Adelchi Azzalini
Dipartimento di Scienze Statistiche
Università di Padova
Padova, Italy
E-mail: azzalini@stat.unipd.it

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 46, Issue 9 *Submitted:* 2010-06-28
February 2012 *Accepted:* 2011-11-30