Introduction to `lnMLE`: An R Package for Marginally Specified Logistic-Normal Models for Longitudinal Binary Data

Bryan A. Comstock and Patrick J. Heagerty
Department of Biostatistics
University of Washington

1 Introduction

Marginal models and generalized linear mixed models are two popular regression approaches for analyzing longitudinal data. Generalized linear mixed models account for within-subject dependence by assuming unobserved random effects. Mean models are then constructed conditional on latent variables and dependence parameters are given by variance components of the random effects. Regression parameters of a nonlinear mixed model have simple interpretation for covariates that vary within a subject, or are time dependent. For covariates that do not vary within a subject, interpretation of the subject-specific covariates may be misleading since they measure a contrast that is not directly observable.

Marginal regression models have been recommended for longitudinal data with time-independent covariates. Marginal regression coefficients describe how the average response changes across subsets of covariate values. The marginal means may be interpreted as averaging over both measurement error and random between-individual heterogeneity.

Heagerty (1999) proposed a hierarchical logistic-normal model which specifies a regression model for both the marginal mean and the random effects variance parameters. *Marginally specified logistic normal models* build regression structure for the marginal mean while allowing valid interpretation of both time-dependent and time-independent covariates. Thus, the proposed likelihood-based approach allows for both estimation of the marginal mean and individual-level effects. The *R*-library `lnMLE` contains the function `logit.normal.mle` that fits the models described in Heagerty (1999). This document is intended to serve as a supplementary, more detailed resource to the `lnMLE` library R help files.

2 General Framework and Notation

We restrict our focus to longitudinal binary response data $\boldsymbol{Y_i} = (Y_{i1},\ldots,Y_{in_i})$ observed on subjects $i = 1,\ldots,N$ at times $\boldsymbol{t_i} = \boldsymbol{t_{i1}}, \boldsymbol{t_{i2}} \ldots, \boldsymbol{t_{in_i}}$. We also assume that there are p exogenous and possibly time-varying covariates $\boldsymbol{X_{ij}} = (\boldsymbol{X_{ij1}}, \boldsymbol{X_{ij2}}, \ldots, \boldsymbol{X_{ijp}})$ recorded for each subject i at each timepoint j.

Our statistical objective is to obtain estimates for the regression of $\boldsymbol{Y_{it}}$ on $\boldsymbol{X_{it}}$. We assume that the regression model properly specifies the full covariate conditional mean defined as $\boldsymbol{\mu_{it}^M} = \mathrm{E}(\boldsymbol{Y_{it}} \mid \boldsymbol{X_{it}}) = \mathrm{E}(\boldsymbol{Y_{it}} \mid \boldsymbol{X_{i1}}, \ldots, \boldsymbol{X_{in_i}})$. This condition assumes that stochastic time-varying covariates are properly modeled through $\boldsymbol{X_{it}}$ and that the current values of the response vector are not good predictors of future covariates. Finally, the marginal generalized linear model specifies $g(\boldsymbol{\mu_{it}^M}) = \boldsymbol{X_{it}}\,\boldsymbol{\beta}$, where $g(\ )$ is a link function and $\boldsymbol{\beta}$ measures the influence of covariates on the average response. In the next sections, we describe the additional assumptions regarding the dependence among the response variables.

2.1 Model Specification

The marginally specified logistic-normal model is specified as a pair of regression models. The first model is a marginal logistic regression for the average response as a function of covariates:

$$\boldsymbol{logit E(Y_{ij} \mid \mathbf{X_{ij}}) = \mathbf{X_{ij}}\beta.} \tag{1}$$

Serial dependence is then modeled by conditioning on a latent variable instead of other response variables:

$$\boldsymbol{logit E(Y_{ij} \mid \mathbf{b_i}, \mathbf{X_i}) = \Delta_{ij} + b_{ij}.} \tag{2}$$

The response vector $\boldsymbol{Y_i}$ is assumed to be conditionally independent given the random effects $\boldsymbol{b_i} = \mathrm{vec}(\boldsymbol{b_{i1}}, \boldsymbol{b_{i2}},\ldots,\boldsymbol{b_{in_i}})$. Finally, we assume that the subject level random effects are normally distributed:

$$\boldsymbol{b_i} \mid \mathbf{X_i} = N(\mathbf{0}, \mathbf{D_i}). \tag{3}$$

The covariance matrix $\boldsymbol{D_i}$ is obtained as a function of the observed times $\boldsymbol{t_i}$ and a parameter vector $\boldsymbol{\alpha}$. The parameter vector $\boldsymbol{\alpha}$ measures the magnitude of variation in the log odds between individuals within groups defined by $\boldsymbol{Z_i}$, where $\boldsymbol{Z_i}$ is a subset of covariates $\boldsymbol{X_i}$. The complete parameter for the marginally specified logistic-normal model is therefore given by $(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

3 lnMLE Implementation in R

The lnMLE R library contains the function logit.normal.mle that fits marginally specified logistic-normal models. The function print.logit.normal.mle is also contained in lnMLE to display a summary of the model output.

This software is for a univariate random effect, $\boldsymbol{b_i}$. In particular, we assume that the random effect $\boldsymbol{b_i}$ is a scaled univariate normal random effect: $\boldsymbol{b_i} = \boldsymbol{\sigma(z_i)} \cdot \boldsymbol{e_i}$, where $\boldsymbol{e_i} \sim \boldsymbol{N(0,1)}$. Using this formulation allows the standard deviation of random intercepts to possibly depend on cluster-level covariates, $\boldsymbol{z_i}$. To structure this relationship we adopt a regression structure for the log of the standard deviation: $\boldsymbol{\log(\sigma(z_i)) = z_i\alpha}$.

3.1 R Function Description: logit.normal.mle

A marginally specified logistic-normal model is called with the following R syntax:

```
logit.normal.mle(meanmodel, logSigma, id = id, n=NULL, alpha=NULL,
beta=NULL, model="marginal", lambda=0.0, r = 20, maxits=50, tol = 1e-3,
data = sys.frame(sys.parent()) )
```

meanmodel	a symbolic description of the mean model to be fit that generally takes the form y $\sim$ x where y are serial binary outcome data and x are the covariates. The covariates x are a series of terms separated by + which specify the linear predictor for y.
logSigma	covariates used to estimate the dependence of $\boldsymbol{\sigma}$ on covariates z. In general, random has the form $\sim$ z where z is a subset of covariates x and is a series of terms separated by +.
id	a vector that identifies the clusters which correspond to the binary response vector given by y.
alpha	(optional) initial parameter estimate(s) of how the covariates z influence the log variance component parameter $\boldsymbol{\sigma}$. The number of estimates provided in alpha should correspond to the number of covariates in z, including an intercept.
beta	(optional) initial estimate(s) of the logistic regression of y on covariate(s) x. The number of estimates provided in beta should correspond to the number of covariates in x.
model	"marginal" for the marginalized model, "conditional" for classic GLMM.
lambda	A likelihood penalty parameter (≥ 0) for alpha. The default is 0.
r	Number of Gauss-Hermite quadrature points. The user may choose r = 3, 5, 10, 20, or 50. The default is 20.

`maxits`	Maximum number of iterations for convergence. The default is 50.
`tol`	tolerance is a measure used in the numerical calculations to determine whether or not convergence of the point estimates has occurred. The default is 1e-3.
`data`	(optional) a data frame containing the variables in the model. If not found in `data`, the variables are taken from `environment(formula)`.

3.2 `logit.normal.mle` Function Output

`Mean parameters (beta):`	
	beta parameter estimates (`estimate`),
	parameter standard errors (`std. err.`)
	z-score test statistics (`Z`)
`Variance Components (alpha):`	
	alpha parameter estimates (`estimate`),
	parameter standard errors (`std. err.`)
	z-score test statistics (`Z`)
`Maximized log-Likelihood of model`	

3.3 An Example

The Madras longitudinal schizophrenia study (Thara et al., 1994) followed 90 first-episode schizophrenics for 10 years with the primary objective of characterizing the natural history of the disease process. The Madras data are included in the `lnMLE` library and may be loaded in R with `data(madras)`.

The data contain serial binary outcome measures $\boldsymbol{y_{it}}$ that denote the presence of positive psychiatric symptoms over the course of $t = 0,\ldots, 11$ months during the first year following hospitalization for schizophrenia for patients $i = 1,\ldots,86$ (denoted by `id`). The dataset also contains the binary indicator of whether or not the patient's `age` at hospitalization $<$ 20 ($0 =$ `age` ≥ 20, $1 =$ `age` < 20), and `gender` ($0 =$ male, $1 =$ female), and the interactions between both of these covariates with time (`month`).

The goal in this example is to illustrate the use of logistic-normal models to characterize the relationship between schizophrenic symptoms and covariates. We are interested in exploring whether the rate of decline in symptoms differs across gender and age-at-onset subgroups.

The following R code is used to examine these data with the R function `logit.normal.mle`:

```
## Load lnMLE library and madras data
library(lnLME)

## logistic-normal model of schizophrenic symptoms:
data(madras)
```

```
attach(madras)

## model symptoms using gender, age, time, and time*age:
model1 <- logit.normal.mle(meanmodel= y ~ gender+month+age+monthXage,
logSigma= ~ 1 + age, id=id, model="marginal", data=madras)
print.logit.normal.mle(model1)

## Example output:
>  ML Estimation for Logistic-Normal Models

>  model = marginal
>  options:     lambda = 0
>                    r = 20

>  Mean Parameters:

>               estimate std. err.    Z
> (Intercept)   0.7783   0.25558  3.045
> gender       -0.7604   0.34778 -2.186
> month        -0.2925   0.03520 -8.309
> age           0.8556   0.41596  2.057
> monthXage    -0.1193   0.07053 -1.691


>  Variance Components:

>               estimate std. err.   Z
> (Intercept)   0.6982    0.1594 4.381
> age           0.3055    0.2573 1.187


>  Maximized logL = -368.5926
```

The $\boldsymbol{\beta}$ parameters are interpreted as comparing the log-odds of schizophrenic symptoms for groups defined by the covariates under the `Mean Parameters` output. In this example, we observe that females were less likely to exhibit schizophrenic symptoms than males at any point during the study (OR: exp(-0.7604) = 0.47). While study subjects who were younger than 20 were more likely to exhibit schizophrenic symptoms at baseline (`month`=0) than those older than 20 (OR: exp(0.8556) = 2.35), younger subjects had a larger rate of decline in the presence of symptoms (OR: exp(-0.2925 -0.1193) = 0.66) than older subjects

(OR: exp(-0.2925) = 0.75) for each additional month during the study period.

We allowed the heterogeneity parameter to depend on age at onset: $\log(\boldsymbol{\sigma_i}) = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1}$age (found under the `Variance Components` output). For subjects over 20 years-old, we observed coefficients of $\boldsymbol{\alpha_0} = 0.6982$ indicating that there is large subject-to-subject variation in the odds of schizophrenic symptoms between subjects (OR: exp(0.6982) = 2.01). There is some evidence for even greater variation in the odds of symptoms between subjects whom are younger than 20 (OR: exp(0.6982 + 0.3055) = 2.73), though this increase was not statistically significant (z=1.187).

4 Conclusions

In this document, we introduced a new R-function `logit.normal.mle` to fit logistic-normal model models to longitudinal binary data and presented an example showing its usage. Marginally specified logistic-normal models adopt a regression structure for the marginal mean instead of the conditional mean. A marginal regression model allows for valid coefficient interpretation with time-independent covariates such as age or gender. Conditional model coefficients in this case may be potentially misleading when subject-specific coefficients are held fixed or controlled for. Finally, the logistic-normal model presented here and in Heagerty (1999) provide valid results with clusters of varying sizes, as when data are missing at random (MAR).

References

[1] Patrick J. Heagerty. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, (55):688–698, 1999.

[2] R. Thara, M. Henrietta, A. Joseph, S. Rajkumar, and W. Eaton. Ten year course of schizophrenia - the madras longitudinal study. *Acta Psychiatrica Scandinavica*, (90):329–336, 1994.