

Subject-specific odds ratios in binomial GLMMs with continuous response

Rapporti di quote specifici per soggetto nei GLMM binomiali a risposta continua

Mariangela Sciandra¹, Vito M.R. Muggeo¹, Gianfranco Lovison¹

Dipartimento di Scienze Statistiche e Matematiche ‘S. Vianelli’, Università di Palermo
e-mail: vmuggeo@dssm.unipa.it

Riassunto: In un contesto di regressione, la dicotomizzazione di una variabile risposta continua è solitamente dettata dall’esigenza di ottenere una misura del rapporto di quote (*odds ratio*) che ne quantifichi l’associazione con uno o più fattori di rischio. Prendendo spunto da un lavoro recentemente pubblicato, in questo articolo si esplora la possibilità di ottenere stime del rapporto di quote da un modello di regressione lineare senza ricorrere alla discretizzazione della variabile risposta, in presenza di dati raggruppati quando un approccio basato su un modello ad effetti casuali viene utilizzato.

Keywords: odds ratio, random-effects, logistic regression, dichotomizing, efficiency.

1. Background

Given $i = 1, \dots, n$ units, for the i subject let Y_i the continuous outcome variable assumed to be dependent on the p -dimensional vector of risk factors $\mathbf{x}_i$; regression models are aimed to investigate the relationships between the response Y and the covariates $X_1, \dots, X_p$. Due to historical reasons and to a presumed superiority in terms of interpretability, the *odds ratio* (OR) is often the preferred association measure used to communicate results in medical and biological studies; therefore even with continuous outcome variables, it is common practice to analyze data through a logistic regression model for the dichotomized response variable $Y^* = I(Y > c)$ for some threshold value c . Often such a cut-off is selected on the basis of biological/medical grounds or to assure a pre-specified number of ‘events’ (for instance c may be the median or any sample quantile of Y); whatever value c , the resulting fitted logistic model provides estimates of the odds ratios which greatly facilitate communication in the medical community. Recently, Moser and Coobs (2004) discuss that with pure continuous response variable Y , the OR may be estimated properly transforming the regression parameters from a continuous regression model of Y versus $X_1, \dots, X_p$ leading to an estimator of OR with greater efficiency than that coming from the dichotomized approach. The paper by Moser and Coobs (2004) deals with independent observations, while in this paper we study at which extent it is possible to generalize the results to account for non-independent observations. We assume that the mixed model framework is employed, where the random coefficients are used to model the subject-specific effects while accounting for the serial correlation among the observations. Section 2 is concerned with statistical methods, where the linear and the binomial generalized linear mixed models are contrasted; Section 3 illustrates results from a small simulation study and Section 4 includes a discussion and some final remarks.

2. The logistic random-effects model for dichotomized response vs. linear mixed models for a continuous response

Random effect models represent one way to handle subject-specific parameters in a model for repeated measurement in which the main interest is in describing the evolution of each subject separately. This approach is based on the assumption that, for every subject, the response can be modelled by a linear (or generalized linear) regression model, but with subject-specific regression coefficients. The Generalized Linear Mixed Models (GLMM) is the most frequently used random-effects model for discrete outcomes. A general framework for a GLMM is the following. Let Y_{ij} be the j -th measurement ($j = 1, \dots, n_i$) for the i -th cluster ($i = 1, \dots, m$). Denoting by $\mathbf{b}_i$ a q -dimensional vector of random effects, assumed to be drawn independently from $MVN(\mathbf{0}, \mathbf{D})$, we assume that conditionally on random effects $\mathbf{b}_i$ the elements Y_{ij} are independent following a generalized linear model but with the linear predictor extended to include subject-specific regression parameters $\mathbf{b}_i$. More specifically, it is assumed that all Y_{ij} have densities from the exponential family and the mean μ_{ij} is modelled through a linear predictor containing fixed regression parameter β as well as subject-specific parameters $\mathbf{b}_i$, i.e., $g(E[Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i]) = g(\mu_{ij}) = \mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i$ for a known link function $g(\cdot)$, and for $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ two vectors containing known covariates' values. In order to obtain OR estimates, the logistic random-effect regression model for the dichotomized response values $Y_{ij}^* = I(Y_{ij} > c)$ is a common choice for analysis of multilevel dichotomous data. It assumes that $Y_{ij}^*|\mathbf{b}_i \sim \text{Bernoulli}(\pi_{ij})$ with $P(Y_{ij} > c|\mathbf{b}_i, \mathbf{x}_{ij}) = P(Y_{ij}^* = 1|\mathbf{b}_i, \mathbf{x}_{ij}) = E[Y_{ij}^*] = \pi_{ij}$ and:

$$\pi_{ij} = (1 + \exp\{-(\mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i)\})^{-1} \quad (1)$$

where the marginal likelihood for this model is obtained integrating over the random effects

$$\prod_{i=1}^m \int \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}^*} (1 - \pi_{ij})^{1-y_{ij}^*} f(\mathbf{b}_i) d\mathbf{b}_i$$

In general, no closed forms are available to solve the above integral because the presence of m integrals over the q -dimensional random effects $\mathbf{b}_i$, therefore numerical integration methods are needed (Crouch and Spiegelman, 1990).

It is well-known that in a binomial GLMM the parameters have a different interpretation than in the corresponding marginal model where the main interest is in the population-averaged evolutions. Fixed effects, infact, have not a marginal interpretation since in general

$$\int \pi_{ij} f(\mathbf{b}_i) d\mathbf{b}_i \neq g^{-1}(\mathbf{x}_{ij}'\beta) = (1 + \exp\{-(\mathbf{x}_{ij}'\beta)\})^{-1}$$

So, in a logistic regression random-effect model the OR interpretation is different from the ordinary logistic regression model because now the OR is a random variable rather than a fixed parameter and this should be led to a different interpretation of the model. However Larsen *et al.* (2000) discuss that the fixed estimates have a nice interpretation in terms of median of the random effects distribution. Therefore focusing on fixed effects estimate

of the OR is meaningful even in a mixed model framework. Rather than to obtain OR estimates from the logistic random effect model for Y_{ij}^* , extension of the idea of Moser and Coobs (2004) leads to consider the LMM for the true continuous response,

$$E[Y_{ij}] = \mathbf{x}_{ij}'\boldsymbol{\alpha} + \mathbf{z}_{ij}'\mathbf{a}_i \quad (2)$$

where now the error terms are independent and identically distributed logistic random variables with mean 0 and variance σ^2 . Model (2) can be used to gain OR estimates of the model (1); simple algebra show that the OR estimates can be obtained by means of $\exp\{\pi\hat{\alpha}_j/(\sqrt{3}\hat{\sigma})\}$ where $\hat{\alpha}_j$ and $\hat{\sigma}$ are estimates from the fitted linear mixed model (2); preserving the continuous nature of Y is expected to lead to a substantial improvement in the OR estimation in terms of unbiased and mean square error of the estimator and consequent width of the estimated confidence intervals.

3. Simulations

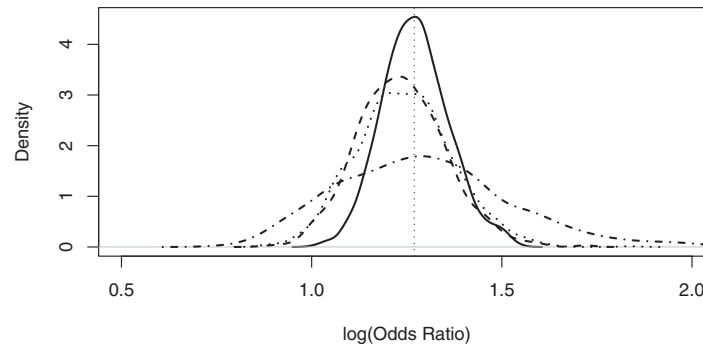
To assess the performance of the aforementioned approach with respect to the dichotomizing procedure, we carried out a few simulations: Table 1 shows results from the simulated model $y_i = b_{0i} + 0.7x_i + \epsilon_i$ where $x_i \sim U(-2, 2)$, $b_{0i} \sim N(0, \sigma_{b0})$ and ϵ_i coming from a logistic distribution with zero mean and scale parameter equal to $\sqrt{3}\sigma/\pi$, where σ^2 is the variance set to one. Three values of σ_{b0} were explored. For each of the one thousand replicates, four models were estimated: i) a random intercept linear mixed model for the response y and explanatory variable x and ii) three random intercept binomial GLMMs for the binary response $I(y > c)$ where c was selected each time to obtain 0.50, 0.30 and 0.10 percent of events. For binomial GLMMs a penalized quasi likelihood approach has been employed, while for LMM we have assumed a Gaussian distribution for the response where the log likelihood has a closed form expression.

Table 1: Monte Carlo estimates of mean and mean square error of the logOR estimator under different models (true logOR ≈ 1.270): LMM and binomial GLMM for the dichotomized response with different percentage π of successes.

σ_{b0}			LMM	GLMM(π)		
				0.5	0.3	0.1
0.3	mean		1.273	1.236	1.244	1.295
	mse		0.078	0.148	0.166	0.516
1.0	mean		1.275	1.201	1.218	1.314
	mse		0.085	0.238	0.274	0.769
2.0	mean		1.273	1.222	1.243	1.308
	mse		0.081	0.354	0.399	1.368

Results highlight how the LMM approach yields estimators with negligible unbiased and mean square errors (mse) which are substantially independent of the variance of the random intercept σ_{b0} ; on the other hand, the logOR estimators via the dichotomizing approach have greater mse which are inflated when the heterogeneity of the intercepts increases and the the number of events decreases. Figure 1 displays simulation results for $\sigma_{b0} = 0.3$.

Figure 1: Smoothed Monte Carlo estimates (based on 1000 replications) of the sampling distributions of the logOR for different models: LMM (continuous line), GLMM(0.5) (dashed line); GLMM(0.3) (dotted line); GLMM(0.1) (dashed-dotted line)



4. Final remarks

When the response variable is continuous and a mixed model framework is employed, odds ratios estimates for the risk factors in the model may be obtained directly from a linear model for the original continuous variable. The loss of information deriving from dichotomizing appears to be substantial, and it is emphasized when random effects are included in the model and in particular when the relevant variance $\sigma_{b_0}^2$ increases. A LMM approach allows to obtain ‘better’ estimates of the OR and furthermore allows to avoid the well-known computational problems which usually complicate the estimation of logistic random effects models. Instead we met convergence problems in fitting binomial GLMMs for the dichotomized variable in some simulation scenarios; such estimating problems are expected to increase when additional random effects (e.g. random slopes) are included in the model, making model estimation really demanding. On the other hand, the linear mixed model does not suffer from such computational burdens since the closed form expression of the log-likelihood. These results can be relevant in practice, since both dichotomizing of continuous outcomes and mixed modelling for longitudinal studies are becoming integral part of medical and biological research.

References

- Larsen K., Petersen, J. H., B.-Jorgesen, E., and Endahl, L. (2000) Interpreting parameters in the logistic regression model with random effects. *Biometrics*, 56, 909–914
- Moser, B. K. and Coobs, L.P. (2004) Odds ratios for a continuous outcome variable without dichotomizing. *Statistics in Medicine*, 23, 1843–1860.
- Stratelli, R., Laird, N., and Ware, J.H. (1984) Random effects models for serial observation with binary response. *Biometrics*, 40, 961–971.
- Crouch, A. C. and Spiegelman, E. (1990) The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464–469.