ANALYSIS OF CORRELATED DISCRETE REPEATED OBSERVATIONS, BIAS IN ESTIMATING REGRESSION PARAMETERS

Schukken YH[1], Gröhn YT[1], McDermott JJ[2]

[1]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca NY 14853, USA
[2]Department of Population Medicine, College of Veterinary Medicine, University of Guelph, Guelph Ontario N1G2W1, Canada

In a previous paper we have addressed the analysis of continuous, Normally distributed correlated data[2]. An important property of such data is the estimation of mean and variance in two unrelated parameters. Since correlation mostly influences the estimation of variance, the mean (or linear predictor in linear models) is not effected by correlation in the data. In the case of discrete data, mean and variance are related, and often estimated using a single parameter. In the case of Poisson distributed data, mean and variance are assumed to be the same and are estimated from a single parameter. In the case of the Binomial distribution the variance is a function of the mean, and both are estimated from a single parameter (assuming n, the sample size, is known). Hence, correct statistical treatment of discrete correlated data is of importance not only for variance estimation, but also for estimation of the mean (linear predictor in GLM).

There are several methods available to estimate regression parameters taking simultaneously into account the extra dispersion that is in the data. These methods include Generalized Linear Mixed Models, where a quasi-likelihood approach is used to estimate variance components and regression components, General Estimation Equations, where regression estimates and variance components are estimated separately, and update each other until convergence is reached, and a true non-linear mixed model estimation procedures where a true likelihood solution is used to obtain simultaneous estimates for variance and regression parameters. In this presentation we aim to contrast the estimation procedures using a data set with correlated binary data.

Materials and Methods

The data were from a trial on cure of intra-mammary infections during the dry period after a single or a double dry cow treatment. These data have been reported before[4]. A total of 110 cows on 12 farms were included in the study. Only quarters infected before dry-off were eligible for cure, making the number of eligible quarters 216. The number of infected quarters per cow ranged from 1 to 4. Essentially, the correlation pattern in the data consists of quarters that are correlated within cow and cows that are correlated within herd, hence a nested multi-level correlation. Since the data are

from only 12 farms, the farms were specifically chosen to yield a high number of cases, and the number of cases per farm was relatively small, we decided to enter herd as a fixed effect into the model, and only estimate a correlation or variance component for quarters within cow. The interest was in estimating the treatment effect, consider age and cell count as possible confounders and evaluate possible treatment X cell count interaction and treatment X age interactions.

The regression model was a basic logistic regression model with a binomial error distribution. Four different methods for estimating parameters in the linear predictor of the regression model were considered:

1. simple logistic regression model, ignoring additional correlation within cow,
2. Generalized Estimation Equation (GEE) to allow for within cow correlation,
3. Generalized Linear Mixed model with a linear random effect to allows for within cow correlation (GLMM macro in Proc Mixed),
4. Non-linear mixed model with a non-linear random effect to allow for within cow correlation (Proc NLMixed).

All analysis were done in SAS v7.0 on a PC. Model 1 and 2 were run in Proc Genmod, model 3 was run with the GLMM Macro that was obtained through the SAS website (www.sas.com), and the experimental Proc NLMixed was used to fit model 4.

Results

All models converged, and yielded results that are tabulated in Table 1. The estimates for model 4, the non-linear mixed model depended somewhat on the initial seed for the random effects parameter. A grid of values was used, and the model with the lowest deviance is reported in this table. The model fit was best in the non-linear mixed model situation as judged by the deviance parameter. The fit of the final model is presented in Figure 1.

Parameters	Logistic	GEE	GLMM macro in Proc Mixed	Proc NLMixed
Intercept	2.73 (.57)	2.79 (.62)	3.84 (.80)	3.77 (1.07)
Age	-.173 (.068)	-.165 (.071)	-.197 (.102)	-.214 (.11)
SCC	-.000063 (.00016)	-.00010 (.00014)	-.00027 (.00019)	-.00018 (.0002)
TRT	.826 (.59)	.789 (.65)	.851 (.792)	1.044 (.8335)
SCC * TRT	-.0015 (.00050)	-.0015 (.00054)	-.0018 (.0006)	-.0020 (.0008)
Deviance	184.6	184.6	215.5 (scaled)	179.3
Variance component	none	r = .205	4.21	2.279 (1.95)

Table 1. Parameter estimates and standard errors for the four estimation methods.

Discussion

The results in table 1 indicate that the simple logistic regression yield the smallest standard error estimates. Since there is considerable correlation left in the data when assuming independence, this is probably incorrect. The estimates of standard error are largest in the results from the true non-linear random effects estimates in Proc NLMixed. The parameter estimates for the random effects models (both GLMM and NLMixed) are always larger then the ordinary logistic regression estimates (unlike the GEE procedure). This is a known feature of random-effect models[3], and the difference between these estimates and the population average estimates of GEE and ordinary logistic regression is a function of estimated variance component (for cow)[1].

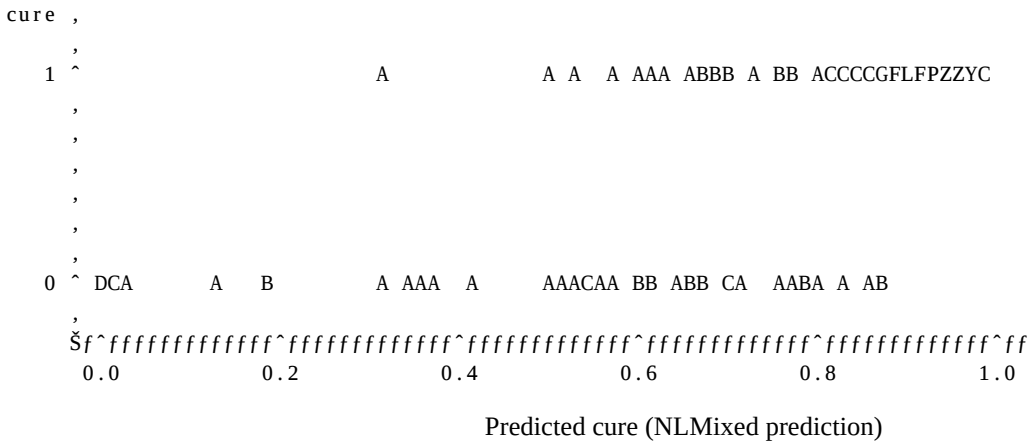

Figure 1. Plot of observed cure versus predicted cure from the NLMixed procedure. Legend: A = 1 observation, B = 2 observations etc., a total of 216 observations.

The models fit by Proc NLMixed can be viewed as generalizations of the random coefficient models fit by the GLMM macro using the Mixed procedure. This generalization allows the random coefficients to enter the model nonlinearly, whereas in Proc Mixed they enter linearly. Also, Proc Mixed assumes the data to be normally distributed after transformation, whereas Proc NLMixed enables analysis of data that are binomial.

References

1. Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. Oxford Science Publications, Oxford. 1994.
2. Grohn YT, McDermott JJ, Schukken YH, Hertl JA, Eicker SW. Analysis of correlated continuous repeated observations: modelling the effect of ketosis on milk yield in dairy cows. Prev. Vet. Med. 1999; 39:137-153.
3. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. Am J Epidemiol 1998; 147:7 694-703.
4. Leslie KE, Bateman K, Barnum D, Schukken YH. Effect estimation of repeated versus single dry cow treatment. Kenya Vet. 1994; 18:149-151.