Open Journal of Statistics, 2013, 3, 1-6
http://dx.doi.org/10.4236/ojs.2013.34A001 Published Online August 2013 (http://www.scirp.org/journal/ojs)

A Comparison of Statistical Methods for Analyzing Discrete Hierarchical Data: A Case Study of Family Data on Alcohol Abuse

Yuanyuan Liang[1], Keumhee Chough Carriere[2*]
[1]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, USA
[2]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada
Email: liangy@uthscsa.edu, *kccarrie@ualberta.ca

Received June 11, 2013; revised July 11, 2013; accepted July 19, 2013

ABSTRACT

Although hierarchical correlated data are increasingly available and are being used in evidence-based medical practices and health policy decision making, there is a lack of information about the strengths and weaknesses of the methods of analysis with such data. In this paper, we describe the use of hierarchical data in a family study of alcohol abuse conducted in Edmonton, Canada, that attempted to determine whether alcohol abuse in probands is associated with abuse in their first-degree relatives. We review three methods of analyzing discrete hierarchical data to account for correlations among the relatives. We conclude that the best analytic choice for typical correlated discrete hierarchical data is by non-linear mixed effects modeling using a likelihood-based approach or multilevel (hierarchical) modeling using a quasi-likelihood approach, especially when dealing with heterogeneous patient data.

Keywords: Non-Linear Mixed Effects Model; Multilevel Model; Generalized Estimating Equations; Mantel-Haenszel Odds Ratio; Specificity; Sensitivity

1. Introduction

The purpose of this paper is to investigate best methodological approaches that frequently arise in the analysis of non-independent discrete hierarchical medical data. There are various methods of handling such types of data. The most general method is a non-linear mixed effects model, which uses a likelihood-based approach. Another method is fitting a multilevel model based on the quasi-likelihood approach proposed by Goldstein (1991) [1]. The generalized estimating equations (GEE) method proposed by Liang and Zeger (1986) uses the concept of quasi-likelihood to fit a generalized linear model (GLM) to clustered data for marginal model building [2]. We compare these methods for their performance, as applied to hierarchical alcoholism data obtained from Edmonton, Alberta, Canada. We outline the strengths and weaknesses of each method.

Data were obtained from a population-based study of mental disorders conducted in Edmonton, Canada. For details of the study design, see Newman and Bland (2006) and the references contained therein [3]. Interviews were conducted with 924 index subjects, called probands, randomly sampled from the population, and 2387 of their first-degree relatives (briefly, relatives). Mental disorders were diagnosed on a lifetime basis (that is, present at the time of interview or ever in the past) using a validated and structured questionnaire. The response variable in this case is the diagnosis of alcohol abuse in relatives, which is dichotomous (1 yes, 0 no). Overall, there were 206 (22.3%) and 461 (19.3%) cases of alcohol abuse among the probands and their relatives, respectively.

We are interested in determining whether (a lifetime history of) alcohol abuse in probands is associated with alcohol abuse in relatives, after adjusting for age and sex of probands and relatives. The data exhibit a hierarchical (or clustered) structure to the extent that the relatives of a given proband have more in common than would be expected in a corresponding random sample from the population. The shared characteristics among relatives from a given family result in data hierarchies and clustering

*Corresponding author.

effects that must be incorporated into the statistical analysis—the usual assumption of the independent and identical distribution (i.i.d.) of the variables is not met.

The data consist a large number of families (n = 924). However, unequal and often small family sizes, ranging from 1 to 12 individuals, create challenges and possible complications in statistical data analysis and model checking. We seek to make general recommendations on how to best estimate the parameters and test the goodness-of-fit of modeling such data.

2. Methods

2.1. Preliminary Analyses

The statistical issue in analyzing hierarchical data, as in the study of the familial aggregation of mental disorders, is to sort out and adjust for associations within a cluster/family. That is, we need to understand whether the presence of the disorder in a proband increases the risk of the disorder in a relative. Investigators engaged in this kind of research typically treat the relatives as if they were a retrospective cohort followed from the beginning of the risk period for the mental disorder until either the onset of the disorder or the time of interview, whichever comes first. The mental disorder status of the proband (Yes/No) is the binary independent variable of primary interest. As a preliminary examination, we compute odds ratios, specifically, crude and Mantel-Haenszel odds ratio estimators [4]. In the Mantel-Haenszel analyses, we accounted for the length of time at risk by including the age of the relative as an independent variable. We considered six age groups: 18 - 24, 25 - 34, 35 - 44, 45 - 54, 55 - 64, and 65+.

2.2. Statistical Modelling

Standard logistic regression that assumes all observations are independent was performed to examine the effect of alcohol abuse in a proband on the risk of alcohol abuse in his/her relatives, adjusting for age and sex of the probands and the relatives. We treated males and the 18 - 24 age group as reference categories. First-order interaction terms were also explored.

However, since the observations are not independent, we also considered regression methods of risk factor modelling that explicitly address the correlation structure of the data: 1) non-linear mixed effects (NLME) model, 2) multilevel model, and 3) generalized linear model (GLM) using generalized estimating equations (GEE).

2.2.1. Non-Linear Mixed Effects Model

Let y_{ij} represent the response of the jth relative of the ith proband (*i.e.* from the ith family), where $i = 1, 2, \cdots, 924$, $j = 1, 2, \cdots, n_i$ with n_i ranging from 1 to 12, and $\sum_{i=1}^{924} n_i = 2387$. Since the response data are binary, we $\pi_{ij} = E\left(y_{ij}\right) = P\left(y_{ij} = 1\right)$ represents the expected value of the response variable for the jth relative in the ith family, and the non-linear link function to model the odds as:

$$f\left(\pi_{ij}\right) = \log\left(\pi_{ij} \big/ \left(1 - \pi_{ij}\right)\right) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \varepsilon_{ij}. \quad (1)$$

Model (1) contains a fixed part, $\boldsymbol{X\beta}$, and a random part, $\boldsymbol{Z\gamma}$. Here, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the fixed and random design matrices, respectively, $\boldsymbol{\beta}$ is a vector of unknown fixed effects, $\boldsymbol{\gamma}$ is a vector of unknown random effects, $E\left(\boldsymbol{\gamma}\right) = 0$ and $\mathrm{var}\left(\boldsymbol{\gamma}\right) = \boldsymbol{\Sigma}$, and ε_{ij} is the unknown random error.

When the model (1) has a univariate random effect with $z_{ij} = 1$, we have a special case, known as a random intercept model. Our data have no other available patient level random effects, and therefore, we consider a random intercept model. Then, $\boldsymbol{\gamma}$ is assumed to follow a normal distribution with mean 0 and variance σ_z^2.

Non-linear mixed effects models are fitted by maximizing an approximation to the likelihood integrated over the random effects. There are several integral approximations available. We use the default optimization technique (dual quasi-Newton) and the default integration method (adaptive Gaussian quadrature) in PROC NLMIXED to obtain the parameter estimates of the model [5,6].

2.2.2. Multilevel Model

For the multilevel model, subscript j denotes the level-1 unit (relative) and subscript i denotes the level-2 unit (family). We define the probability π_{ij} as a function of an intercept and several explanatory variables similar to the non-linear mixed effects model considered above.

The full model in terms of π_{ij} can be written as

$$\begin{aligned} \mathrm{logit}\left(\pi_{ij}\right) &= \beta_{0i} + \boldsymbol{X\beta} \\ \beta_{0\mathrm{i}} = \beta_0 + \xi_i &\text{ where } \xi_i \sim N\left(0, \sigma_z^2\right). \end{aligned} \quad (2)$$

Here, ε_{ij} is the variance of the level-1 (relative) random term with mean 0 and variance 1. The intercept is being modeled as random at the level of the family (level-2), that is, the probability of a relative having an alcohol abuse disorder at the reference values (all the explanatory variables set to zero) is different across families, while all the other parameters $\left(\boldsymbol{\beta}\right)$ are fixed. We assume that the random part of the intercept, ξ_i, follows a normal distribution with mean 0 and variance σ_z^2. Note that this model is identical to model (1) under the random intercept model for two level hierarchical data, but it is capable of building multilevel models. Also it uses an algorithm, which is different from that for NLME to analyze the data when using commercial software.

We use the second-order penalized quasi-likelihood (PQL2), which has been shown to be least biased, com-

pared to the first-order marginal quasi-likelihood Method [7]. Bootstrap estimation is an alternative, as it corrects the bias associated with the quasi-likelihood procedures [8]; however, the improved accuracy is usually obtained at the expense of lengthy computational time.

2.2.3. GEE Method

Generalized estimating equations (GEE) method proposed by Liang and Zeger (1986) uses the concept of quasi-likelihood to fit a generalized linear model (GLM) to clustered data. The GEE method for estimating $\boldsymbol{\beta}$ is an extension of the independence estimating equation to the correlated data. The GEE is given by the score function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\pi}_i^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}) \right) = 0 \qquad (3)$$

with K = 924 in our case. If $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\boldsymbol{Y}_i$, then the true covariance matrix of $\boldsymbol{Y}_i$ is given as $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2} / \phi$, where $\boldsymbol{A}_i$ is an $n_i \times n_i$ diagonal matrix with $\mathrm{var}(\boldsymbol{\pi}_{ij})$ as the jth diagonal element, and ϕ is the scale parameter. Liang and Zeger (1986) proposed GEE for $\boldsymbol{\beta}$ based on Equation (3) where ϕ is replaced by an estimator $\hat{\phi}(\boldsymbol{\beta})$, which is $\sqrt{K}$-consistent given $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ is replaced by an estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)$, which is also $\sqrt{K}$-consistent given $\boldsymbol{\beta}$ and ϕ.

The correlation parameters $\boldsymbol{\alpha}$ and the scale parameter ϕ are estimated iteratively using the current value of the parameter vector $\hat{\boldsymbol{\beta}}_r$ at the rth iteration. Finally, we calculate the appropriate functions of the Pearson residuals.

Under certain regularity conditions, the solution to Equation (3) is consistent and asymptotically multivariate normal regardless of whether the working correlation matrix has been modeled correctly. To account for the family effect using GEE in marginal model building, we assume that the correlation among members of the same family is the same for all 924 families. That is, they all share the same working correlation matrix. We analyze the alcoholism data using three different structures for the working correlation matrices—exchangeable, AR(1) and unstructured.

2.3. Model Diagnostics

To check the prediction accuracy achieved under each model, we compare the model prediction results and the observed presence of the disorder using a classification table, where a predicted probability of alcohol abuse of 0.5 or more is classified as a "positive" prediction. For the diagnostic tests, two critical components determine the model's accuracy: sensitivity (the probability that a test is positive given that the person has the disorder) and specificity (the probability that a test is negative given that the person does not have the disorder). In a perfect model, all cases will be on the diagonal of the classification table and the overall percent correct will be 100.

The data were analyzed using commercial software: for the non-linear mixed effects model, we used PROC NLMIXED in SAS (SAS Institute, Cary, NC), for the hierarchical model, we used MLwiN (Centre for Multilevel Modelling, University of Bristol, UK), and for the GLM/GEE model we used PROC GENMOD in SAS. The technical details of these three non-linear modeling approaches and the SAS codes are available upon requests.

3. Results

Table 1 provides summary information about the age and sex distribution of relatives and probands respectively. In a preliminary analysis, we assume that relatives in a given family are independent; that is, they are uncorrelated. In all models considered, we treat the alcohol abuse status of probands as the exposure, and the alcohol abuse status of relatives as the outcome: *Exposure* is defined as Yes if the proband has had an alcohol abuse problem, No if the proband has no alcohol abuse problem; and *Outcome* is defined as Yes if the relative has had an alcohol abuse problem, No if the relative has no alcohol abuse problem.

Table 2 summarizes the contingency table analysis of

Table 1. Age and sex distribution of probands and relatives.

Category		Relatives		Probands	
		Count	%	Count	%
Age	18 - 24	298	12.5	61	6.6
	25 - 34	636	26.6	248	26.8
	35 - 44	463	19.4	181	19.6
	45 - 54	333	14.0	129	14.0
	55 - 64	356	14.9	135	14.6
	65+	301	12.6	170	18.4
Sex	Female	1358	56.9	604	65.4
	Male	1029	43.1	320	34.6

Table 2. Crude and Mantel-Haenszel odds ratio estimates.

Statistic	Crude[a]	MH[b]	MH[c]
Odds ratio	1.525	1.710	2.421
95% CI	(1.208, 1.926)	(1.325, 2.206)	(1.771, 3.309)
Width of CI	0.718	0.881	1.538

[a]Crude odds ratio based on a standard 2X2 table; [b]Mantel-Haenszel odds ratio based on a stratification by age and sex of relatives; [c]Mantel-Haenszel odds ratio based on a stratification by age and sex of relatives and probands.

exposure versus outcome, with and without adjustment. From this table, we can see that alcohol abuse in a proband is strongly associated with that in a relative. The second Mantel-Haenszel odds ratio estimate incorporates stratification by age and sex of both probands and relatives; alcohol abuse in a proband increases the odds of alcohol abuse in a relative by more than two folds.

In regression modelling, we considered interactions in model (1) with all methods, but none of the interaction terms were found to be statistically significant. As mentioned in the Methods section, for two level hierarchy, NLME and hierarchical models were basically identical. Any differences are more of algorithmic and computational in nature than anything structural in modeling. Note that these two models allowed probands to be random, accommodating their natural heterogeneity, while in GEE modeling, we only get the marginal effects.

Table 3 summarizes the estimates, along with their standard errors in the final model. The results are broadly similar in all regression approaches qualitatively. As can be seen, alcohol abuse in a proband significantly increases the odds of alcohol abuse in a relative. The age of the proband is not statistically significant. For the age of relatives, there is a decreasing trend in the odds of alcohol abuse as they get older. For both probands and relatives, sex is highly statistically significant, but with differing effects: for probands, being female increases the odds of alcohol abuse for her relatives, while for relatives the opposite is true. The GEE estimated the within-cluster correlation to be rather small at 0.074, indicating that at the marginal level, the observations were more or less independent. However, both parametric methods (NLME and hierarchical models) estimated the random effects to be quite significant at 0.741 (SE = 0.237) and 0.726 (SE = 0.170), respectively.

Table 3. Parameter estimates based on the regression approaches.

Parameter	Non-linear mixed model	Multilevel model PQL2	GEE
Intercept	−1.079 (0.222)	−1.082 (0.217)	−0.944 (0.218)
Alcohol abuse	0.842 (0.182)	0.844 (0.179)	0.765 (0.169)
Female proband	0.537 (0.166)	0.539 (0.165)	0.480 (0.158)
Female relative	−2.033 (0.147)	−2.025 (0.137)	−1.826 (0.118)
Relative 25 - 34	0.213 (0.200)	0.212 (0.198)	0.184 (0.186)
Relative 35 - 44	−0.026 (0.217)	−0.026 (0.215)	−0.017 (0.199)
Relative 45 - 54	−0.421 (0.242)	−0.421 (0.242)	−0.391 (0.218)
Relative 55 - 64	−0.622 (0.243)	−0.622 (0.243)	−0.559 (0.215)
Relative 65+	−1.405 (0.293)	−1.400 (0.300)	−1.280 (0.257)
Random effect/ correlation	0.741 (0.237)	0.726 (0.170)	0.074[a]

Note: Entries are estimates (standard error). GEE method assumes exchangeable working correlation structure. Relative aged 18 - 24 is the reference category. [a]Parameter in the working correlation matrix, *i.e.* correlation coefficient between relatives related to the same proband. Since GEE considers the correlation among clustered observations as a nuisance, no standard error is calculated.

The prediction accuracy using the diagnostic tests for each approach is given in **Table 4**. All models performed well at predicting negative cases correctly, while there were shortcomings in the prediction of positive cases. The non-linear mixed model was the best overall, followed by the multilevel model. The popular GEE model did not perform as well in its predictions. It should be noted that the sensitivities and specificities reported here are a function of the independent variables included in the model, and these were chosen purely for illustrative purposes.

Finally, **Table 5** compares the adjusted odds ratios and corresponding 95% confidence intervals from the various analyses. It demonstrates that alcohol abuse in a proband more than doubles the odds of alcohol abuse in a relative. It is notable that the three regression methods that take the intra-familial correlations into account produce narrower confidence intervals than the Mantel-Haenszel approach and the standard logistic regression, which treated all observations as independent.

4. Discussion

This paper reviews methods of analyzing hierarchical data in an effort to highlight their weaknesses and strengths and draw general guidelines for their use in

Table 4. Model accuracy.

Method	Sensitivity	Specificity
Non-linear mixed model	27.98	98.65
Multilevel model (PQL2)	25.81	98.65
GEE	9.54	98.18

Note: Entries are in percentages.

Table 5. Summary of odds ratio estimates.

Method	Coef.	OR	95% CI	Width of CI
Standard logistic regression	0.766	2.152	(1.61, 2.87)	1.26
Non-linear mixed model	0.842	2.322	(1.97, 2.68)	0.71
Multilevel model (PQL2)	0.844	2.326	(1.97, 2.68)	0.71
GEE	0.765	2.150	(1.82, 2.48)	0.66
MH Odds Ratio[a]		2.421	(1.61, 2.86)	1.25

[a]Mantel-Haenszel odds ratio is adjusted by age of relatives and sex of both relatives and probands.

analyzing medical and health data with correlated binary responses. To illustrate, we made use of alcohol abuse data from a family study conducted in Edmonton, Canada.

The non-linear mixed effects model assumes that the error distribution is normal, while allowing for the heterogeneity of the data in the form of mixed effects of some covariates. In this situation, multilevel models can be viewed as a special case of non-linear mixed effects models, but they are especially useful when the data have more than two levels of hierarchies. Unlike these two approaches, which assume a correlated binomial distribution of the data, the GEE method does not require the data to follow a particular parametric distribution. However, if the number of clusters is very large, all three methods are expected to perform similarly.

Using the alcohol abuse data, we found that the non-linear mixed model resulted in the best overall model prediction with the additional advantage of requiring only a moderate amount of computing time. However, only a limited number of researches have been done to check the model assumptions [9-11]. Further research is needed to improve this aspect of the non-linear mixed model.

The multilevel method allows for a model with several levels, but for two-level hierarchical data such as was used in this study, it is essentially the same as the usual non-linear mixed effects model. A disadvantage of this approach is that the algorithm may not converge, especially when the cluster size is small with few clusters. Furthermore, the computation time to convergence is relatively long.

In the GEE approach, we choose the exchangeable structure of the working correlation matrix, assuming all relatives of the proband have the same correlation. We note that Liang and Zeger (1986) originally considered the correlation among clustered observations as a nuisance, while the regression parameters are the primary interest [2]. With the GEE approach, regression parameters can be estimated consistently but not necessarily with complete efficiency, whether the working correlation structure is correct or not. This consistency is based on the assumption that the regression parameters and the association parameters are orthogonal to one another, even when they are not [12]. However, the GEE method may still be preferable in some cases, when a population averaged level analysis is suitable for the research questions and objectives. Further, the computational algorithm is relatively fast and it makes weaker assumptions about the structure of the variance-covariance matrix of the response vector.

Regarding the model's prediction accuracy, in general, the model's specificity is quite high, but rather low in sensitivity. In other words, the model has much more difficulty predicting cases that have an alcohol abuse problem, which is a common situation when there are no strong risk factors for modelling. In the absence of random effects, all models have an essentially equal ability to predict the outcome. However, when there are significant random effects, the prediction level improves by properly accounting for the random effects in the model.

5. Conclusion

Overall, the non-linear mixed effects approach to analysis of these data seems quite competitive with the multilevel method in terms of convergence properties. The random family effects were significant, reflecting heterogeneity among families, and both the non-linear mixed effects model and the multilevel model captured them effectively. The popular marginal modeling via the GEE method may still be preferable because of its computational ease and relaxed distribution assumptions. However, caution is advised, as it might underestimate the odds ratio and its standard error, as indicated in this case study.

REFERENCES

[1] H. Goldstein, "Nonlinear Multilevel Models, with an Application to Discrete Response Data," *Biometrika*, Vol. 78, No. 1, 1991, pp. 45-51. doi:10.1093/biomet/78.1.45

[2] K. Y. Liang and S. L. Zeger, "Longitudinal Data-Analysis Using Generalized Linear-Models," *Biometrika*, Vol. 73, No. 1, 1986, pp. 13-22. doi:10.1093/biomet/73.1.13

[3] S. C. Newman and R. C. Bland, "A Population-Based Family Study of DSM-III Generalized Anxiety Disorder," *Psychological Medicine*, Vol. 36, No. 9, 2006, pp. 1275-1281. doi:10.1017/S0033291706007732

[4] N. Mantel and W. Haenszel, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *Journal of the National Cancer Institute*, Vol. 22, No. 4, 1959, pp. 719-748.

[5] J. C. Pinheiro and D. M. Bates, "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model," *Journal of Computational and Graphical Statistics*, Vol. 4, No. 1, 1995, pp. 12-35.

[6] J. Nocedal and S. J. Wright, "Numerical Optimization," Springer-Verlag, New York, 1999. doi:10.1007/b98874

[7] N. E. Breslow and D. G. Clayton, "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, Vol. 88, No. 421, 1993, pp. 9-25.

[8] H. Goldstein and J. Rasbash, "Improved Approximations for Multilevel Models with Binary Responses," *Journal of the Royal Statistical Society Series A—Statistics in Society*, Vol. 159, 1996, pp. 505-513.

[9] Y. Yano, S. L. Beal and L. B. Sheiner, "Evaluating Pharmacokinetic/Pharmacodynamic Models Using the Posterior Predictive Check," *Journal of Pharmacokinetics and Pharmacodynamics*, Vol. 28, No. 2, 2001, pp. 171-192.

doi:10.1023/A:1011555016423

[10] P. J. Williams and E. I. Ette, "Determination of Model Appropriateness," In: H. C. Kimko and S. B. Duffull, Eds., *Simulation for Designing Clinical Trials*: *A Pharmacokinetic-Pharmacodynamic Modeling Prospective*, Marcel Dekker, New York, 2003, pp. 68-96.

[11] F. Mentre and S. Escolano, "Prediction Discrepancies for the Evaluation of Nonlinear Mixed-Effects Models," *Journal of Pharmacokinetics and Pharmacodynamics*, Vol. 33, No. 3, 2006, pp. 345-367. doi:10.1007/s10928-005-0016-4

[12] K. Y. Liang, S. L. Zeger and B. Qaqish, "Multivariate Regression Analysis for Categorical Data," *Journal of the Royal Statistical Society, Series B*, Vol. 54, No. 1, 1992, pp. 3-40.

