A Note on Choosing a Mathematical Model for Repeated Measurements

Ahsène Lanani
Department of Mathematics. Faculty of Exact Sciences.
University Constantine 1. Constantine 25000 Algeria.

Abstract

Our problem consists of choosing the 'best' model among the family models which are frequently used in the field of Biostatistics; Medicine; Biology; Agricultural sciences; etc.... The methods used for the estimation of the model parameters are the maximum likelihood and the resolution of the estimating equations, known as Generating Estimating Equations. After having described the various models and methods as well as the majority of the developments known to us so far, we opted for the mixed linear models with the maximum likelihood method

1. Introduction

The frequency of the repeated measures in biology, epidemiology and the problems of health in general is at the origin of the increasing interest of biostatisticians for the methods of statistical analyses which are adapted specifically to the correlated data. The choice of a statistical model among a family of models and an analysis method among others, is not an easy task. This choice depends on the applicability, the aim, the structure of the sample or on the degree of dependence inside the individual groups.

The longitudinal studies aims at observing any individual on two occasions or more over wide periods, by taking account of time; on the other hand, repeated measurements are taken during one period of study which is very short, by taking account of the experimental conditions (Ware 1985). The book of Diggle et al. 2002 is a complete work treating the longitudinal data analysis. For longitudinal data; the analyses are often concerned with the investigation of changes over time of a characteristic which is repeatedly measured for each study subject or experimental unit. In medical sudies, the measurements might be cholesterol level; serum glucose, lung volume or blood pressure.

Longitudinal data may be called balanced when the same number of time points is available on each unit and time intervals between pairs of corresponding obsrvations are the same for all units; however, obsevations on the same unit need not be equally spaced. In practice, these data are often unbalanced; that is all the individuals are not observed at equally space time points and the obsevation numbers are not equal for the individuals. So, methods based on the standard multivariate linear model are not available.

For measurements in series (repeated measurements), one can, for example, uses the time series; though in practice, the calculative problems repeated on these time series which are generally short and numerous which make these methods inapplicable, in rending the passage to other methods. One can, for example, use the mixed linear models which consist of using all these series at the same time; the method of least squares; the bootstrap; the generalized linear models which often use quasi-likelihood; the marginal models, etc...

This note comments on the choice of a statistical model for longitudinal data or unbalanced repeated measurements analysis. Once a model is chosen, the estimation of its parameters is carried out by a standard method among a large given family, such as the maximum likelihood or the weighted least squares.

2. Models

2.1. Random effects models

The linear regression whose objective is the study of the relation between a variable response (explained variable) and one or more explanatory variables, is based on the linear model (LM). In order to explain variability between the various individuals, random effects were introduced into the explanatory part of the traditional linear models. That gives rise to the mixed linear models or random-effects models which are noted by LMM, or also sometimes by certain authors by L2M.

This first family; namely the mixed linear models are widely used (Harville 1977 ; Laird and Ware 1982 ; Chi and Reinsel 1989; verbeke and Molenberghs 2000; Littell et al. 2000).These models prove to adapt suitably to the longitudinal data and repeated balanced or unbalanced measurements, even in the presence of missing data. However, on the one hand, they suppose that the data follow gaussian distributions ; on the other hand, the calculative problems pose a problem in spite

of considerable developments of software and procedures, such as PROC MIXED or GENMOD of the SAS system (Littell et al. 1996). When one uses the maximum likelihood (ML), the obtained normal equations are generally nonlinear. Consequently, these equations are solved by iterative processes, such as the EM algorithm (Dempster et al.1977; Laird et al.1987); the Newton Raphson algorithm (Lindstrom and Bates 1988); the Fisher scoring algorithm (Jennrich and Schluchter 1986; etc...). To avoid the slowness of certain algorithms and the problems of convergence which is sometimes local rather than global; an alternative consists of switching on noniterative methods, especially for the variance-covariance matrix estimate of the considered model. To achieve this, we propose the following method.

2.2. The weighted least squares method

The weighted least squares method gives unbiased and consistent estimates. However, it does not make valid the tests of the confidence intervals which are based on normality (data are supposed normally distributed). This problem can be solved by the use of the nonparametric tests (Zerbe 1979); but these methods are applied only for balanced data, which is not often the case for longitudinal data or repeated measurements.

2.3. The bootstrap

The Bootstrap (Efron and Gong 1983); is another method to avoid the normality assumption. The idea is to work with an estimator of a sample density. However, there are disadvantages such as heaviness in calculations and the missing data can also pose problems.

2.2. Marginal models

There is another alternative which is different from the preceding ones. The marginal models wich consists of solving the generalized estimating equations (GEE). This method uses, on the one hand, the generalized linear models (GLM) (Mc Cullagh and Nelder 1989) and on the other hand, the generalized estimating equations (Liang and Zeger 1986), which are an extension of quasi-likelihood (QL) (Wedderburn 1974). However, one obtains a rough variance-covariance matrix estimate of the individuals. In addition, the variance is regarded as a nuisance parameter. We are interested much more in the regression parameters. In this GEE method, the true matrix of correlation is replaced by a matrix whose choice is arbitrary, it is a working correlation matrix.

This last method, which was introduced for the first time by Liang and Zeger (1986); is a current controversial problem, as far as its use is concerned; because, ignoring the correlation, affects the inference of the regression coefficients, on the one hand; and on the other hand, the regression coefficients estimates will be inefficient (Crowder 1995;2001).

Of course, for the selection or comparison of models, some criteria, such as the AIC (Akaike Information Criterion) and the BIC (Bayes Information Criterion) do exist, wich we did not mention. We have only outlined a brief description of the various models in a general and not a particular context (without including particular data).

Among these families of models, the most used in quantitative genetics; medicine; biology; ecology; biostatistics, as well as in other fields, are the first (the random effect models) and the last (the marginal models). This is why we insist on the completed work concerning these models.

3. Notes and discussion

Advantages and disadvantages of the marginal models and generalized estimating equations are evoked in several works. One can quote those of Zhao and Prentice (1990); Prentice and Zhao (1991); Liang et al.(1992); Fitzmaurice and Laird (1993); Park (1993); (Crowder 1995);Lindsey and Lambert (1998); Crowder (2001); among others.

Recall that Liang and Zeger (1986) introduced their approach for the analysis of correlated data. Their idea was to model the marginal means of the variable response and to estimate the regression parameters by the resolution of the generalized estimating equations. These equations use a working correlation matrix, which depends on a parameter α. This matrix is arbitrary and can not be correctly specified. The authors proposed thereafter an estimator of the variance regression parameters, known as robust estimator or 'sandwich estimator' and showed that the regression parameter estimates and their variances are convergent even if the working correlation matrix is badly specified.

Prentice (1988); extended this idea in the context of binary responses by introducing estimating equations for the correlation parameter noted by α. The objective was to jointly estimate the parameters of regression and correlation.

Prentice and Zhao (1991) and Zeger and Liang (1992) generalized this method for an unspecified responses.

Through examples taken for the working correlation matrix and for the true correlation matrix, Crowder (1995) showed that the estimator of α can not be consistent (if it does exist at all); this raises a problem on the first assumption of theorem 2 of Liang and

Zeger (1986). Wherby to satisfy this assumption, the situations where the estimator of α is K^{1/2} consistent (K is the individual number) are sought.

Park and Shin (1995) criticized the work of Crowder (1995) and contradicted the results author's by simulations. However, these simulations were made on small size samples (n=25 and n=100). What about large samples then ? taking in consideration that the work of Crowder (1995) concerned large samples which raised controversial over the asymptotic results of Liang and Zeger (1986).

To solve the problem of disadvantages of the generalized estimating equations of Liang and Zeger (1986), Crowder (2001) proposed improvements of those equations by combining a noted approach GE (' Gaussian Estimation' based on the maximum likelihood) with the GEE equations. This method is much more based on the GE method. The author concluded that it is more advantageous and easier to maximize a function, such as the likelihood, and that a maximum almost always exists, even if it is local than to solve equations,for example, the GEE equations, which sometimes can not have solutions.

Other authors tried to make improvements concerning GEE equations.

In particular, Lipsitz et al. (1991) proposed the odds ratio (OR) per pair such as a measure of association within-group instead of the correlation or covariance.

Liang et al. (1992) like Fitzmaurice and Laird (1993) also used the odds ratio.

Comparisons between the approach of the Maximum likelihood and that of GEE equations were done by Park (1993) who went for the first method.

Lindsey and Lambert (1998) underlined the advantages and especially the disadvantages of the marginal models (for example a treatment can be efficient on average whereas it is bad for each subject). However, the authors underlined that these models can be adapted for descriptive studies, such as the epidemiological studies. In fact, these models can be only applied with a great precaution in the experimental studies, such as the clinical trials. Examples are given by authors to compare the marginal models versus the conditional ones.

Hall and Severini (1998) proposed an extension of the GEE in order to improve the effectiveness of estimators of the association parameters α. Their method is entitled extended generalized estimating equations (EGEE method).

Lastly, let us note that Hu and Lachin (2001) insisted on the fact that various working correlation matrices arrive at various conclusions by following a study on the treatment of diabetes.

4. Conclusion

Based oneself on the results of the literature concluded by the various authors and contradicted by others, one can say that the choice of the working correlation matrix, let alone the choice of the GEE method by using marginal models, is rather delicate and that this method remains very debatable; especially, with respect to that of the maximum likelihood in the context of the random effects models. Therfore one notes that the least remains the 'best' method and the adequate model too for analysing longitudinal data or repeated measures balanced and especially unbalanced.

References

[1] Chi E.M. and Reinsel G.C.(1989). Models for longitudinal data with random effects and AR(1) errors. Journal of the American Statistical Association 84, 52-59.

Crowder M.(1995). On the use of a working matrix correlation in using GLM for repeated measures. Biometrika 82, 407-410.

Crowder M.(2001). On repeated measures analysis with misspecified covariance structure. Journal of the Royal Statistical Society, Series B 63, 55-62.

Dempster A.P.; Laird N.M. and Rubin D.B.(1977). Maximum likelihood with incomplete data via the E-M algorithm. Journal of the Royal Statistical Society, Series B 39, 1-38.

Efron B and Gong G.A.(1983). A leisurely look at the bootstrap, the jacknife and cross-validation. The American Statistician 37, 170-174.

Fitzmaurice G.M. and Laird N.M.(1993). A likelihood based method for analysing longitudinal binary responses. Biometrika 80, 141-51.

Hall B. and Severini T.(1998). Extended Generalized Estimating Equations for clustered data. Journal of the American Statistical Association 93, 1365-1374.

Harville D.A.(1977). Maximum likelihood approaches to variance components estimation and to related problem. Journal of the American Statistical Association 72, 320-329.

Hu M. and Lachin J.M.(2001). Application robust estimating equations to the analysis of quantitative longitudinal data.Statistics in Medicine 20, 3411-3428.

Jennrich R.I. and Scluchter M.D.(1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. Biometrics 42, 805-820.

Laird N.M and Ware J.H.(1982). Random effects models for longitudinal data. Biometrics 38, 963-974.

Laird N.M.; Lange N. and Stram D.(1987). Maximum likelihood computations with repeated measures : application of he E-M algorithm. Journal of the American Statistical Association 82, 97-105.

Liang K.Y.and Zeger S. L.(1986). Longitudinal data analysis using GLM. Biometrika 73, 13-22.

Liang K.Y.; Zeger S. L. and Qaquish B.(1992). Multivariate regression analysis for categorial data. Journal of the Royal Statistical Society, Series B 54, 3-40.

Lindsey J.K.and Lambert P.(1998). On the appropriateness of marginal models for Repeated Measurements in Clinical Trials. Statistics in Medicine17, 447-469.
Lindstrom M.J. and Bates D.M.(1988). Newton-Raphson and E-M algorithm for linear mixed effect models for repeated - measures data. Journal of the American Statistical Association 83, 1014-1022.
Littell R.C.;Milliken G.A.;Stroup W.W. and Wolfinger R.D.(1996). System for mixed models. SAS institute inc: Cary,N.C.
Littell R.C.;Pendergast J.and Natarajan R.(2000). Modelling covariance Structure in the analysis of repeated measures data. Statistics in Medicine 19, 1793-1819.
Lipsitz S.R.;Laird N.M. and Harrington D.P.(1991).Generalized estimating equations for correlated binary data : using odds ratio as a measure of association.Biometrika 78, 153-160.
McCullagh P. and Nelder J.A.(1989). Generalized linear model, 2nd edition. London : Chapman and Hall.
Nelder J.A. and Pregibon D.(1987). An extended quasi-likelihood function. Biometrika 74, 221-232.
Park T.(1993). A comparaison of the GEE approach with the maximum likelihood approach for repeated measurements. Statistics in Medicine 12, 1723-1732.
Park T. and Shin D-Y.(1999). On the use of working correlation matrices in the GEE approach for longitudinal data. Communications in Statistics -Simulation 28, 1011-1029.
Prentice R.L.(1988). Correlated binary regression with covariates specific to each binary observation. Biometrics 44, 1033-1048.
Prentice R.L. and Zhao L.P.(1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47, 825-839.
Verbeke G. and Molenberghs G. (2000). Linear Mixed Models for longitudinal data. New York : Springer.
Wedderburn R.W.M.(1974). Quasi-likelihood function; GLM and the Gauss- Newton method. Biometrika 61, 439-447.
Zhao L.P. and Prentice R.L.(1990).Correlated binary regression using a quadratic exponential model. Biometrika 77, 642-648.
Zeger S.L and Liang K.Y.(1992). An overview of methods for the analysis of longitudinal data. Statistics in Medicine 11, 1825-1839.
Zerbe G.O.(1979). Randomization Analysis of Growth Curves. Journal of the American Statistical Association 74, 215-221.