International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-2, Issue -5, September 2015

Modeling Repeated Measures

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

ABSTRACT
During the study of longitudinal data or repeated measures, we are often concerned with the choice of a good mathematical or statistical model to approach reality. In this paper, we present different models. Our goal is the choice of the suitable ones.

Keywords
Modeling data, longitudinal data, repeated measures, variance covariance
.

1. INTRODUCTION
Generally there is no problem to study simple statistical data because there is a wide variety of statistical models; especially linear regression models (Andrew and Jeniffer 2006; Bonneu; Chouquet 2010). However, when the data are longitudinal; repeated; grouped or when there are missing or censored data (Droesbeke et al.1989) in the statistical analysis, the problem which arises is the correlation between these data. This situation is common in medicine; biology; etc… generally in the life sciences. In this situation it is difficult to choose the right model. Among the suitable models, we can take the linear mixed model. 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. When the data are unbalanced; that is all the individuals are not observed at equally space time points and the observation numbers are not equal for the individuals. In this situation, methods based on the standard multivariate linear model are not available. For repeated measurements, we can use the time series; though in practice, the calculative problems repeated on these time series which are generally short and numerous make these methods inapplicable, in rending the passage to other methods. We 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... 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

2.1.1 Introduction
In order to explain variability between the various individuals, random effects were introduced into the explanatory part of the traditional linear model (LM). That gives rise to the mixed linear models or random-effects models which are noted by LMM (or 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; Fitzmaurice et al. 2004).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 normal 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 non iterative methods, especially for the variance-covariance matrix estimate of the considered model.

2.1.2 Variance-Covariance Structures
In this section we present several different covariance structures. The goal is to choose a parsimonious structure.

2.1.2.1 Simple structure
This structure suppose that observations on the same subject are independent.

2.1.2.2 Compound symmetry structure (C.S)
In this situation observations on the same subject have homogeneous covariance and variances too.

2.1.2.3 AR(1)Structure
For this structure, the variances are homogeneous and the correlation decrease toward zero when the lag increase

2.1.2.4 AR(1)RE Structure
In this situation, we have an autoregressive structure with random effect for the subject.

2.1.2.5 Unstructured
There is no special structure in this case. It is a general case.

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. This method is useful when the theoretical distribution of a statistic of interest is complicated or unknown, or when the sample size is insufficient for statistical inference. 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.

2.4. Marginal models
The marginal models consist 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, which 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; ecology; engineering, 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 DISCUSSIONS
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 cannot 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 cannot be consistent (if it does exist at all); this raises a problem on the first assumption of theorem 2 of Liang and Zeger (1986). Whereby 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 cannot 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 those 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. CONCLUSIONS

In this paper, we present different models. Our goal is the choice of the suitable ones. Based oneself on the results of the literature concluded by the various authors and contradicted by others, We 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. Therefore we note that the least remains the best method and the adequate model too for analysing longitudinal data or repeated measures.

REFERENCES

[1] Andrew, G. and Jeniffer, H. 2006.Data analysis using regression and multilevel hierarchical models. *Cambridge University Press.*

[2] Bonneu, M. and Leconte, E. Modèle linéaire. *Poly. Cours Stat. Eco. www.info.univ-angers.fr.*

[3] Chi E.M. and Reinsel G.C. 1989. Models for longitudinal data with random effects and AR(1) errors. *J. Amer. Stat. Assoc. 84, 52-59.*

[4] Chouquet, C. 2006. Modèles linéaires. *Labo. Stat. Proba. U. P. Sabatier Toulouse.*

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

[6] Crowder, M. 2001.On repeated measures analysis with misspecified covariance structure. *J. Roy. Stat. Soc., Series B 63, 55-62.*

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977.Maximum likelihood with incomplete data via the E-M algorithm. *J. Roy. Stat. Soc., Series B 39, 1-38.*

[8] Droesbeke, J.J., Fichet, B. and Tassi, Ph. 1989. Analyse statistique de durée de vie. Modélisation de données censures. *Economica.*

[9] Efron, B. and Gong, G.A. 1983.A leisurely look at the bootstrap, the jacknife and cross-validation. *Amer. Stat. 37, 170-174.*

[10] Fitzmaurice, G.M. and Laird, N.M. 1993. A likelihood based method for analysing longitudinal binary responses. *Biometrika 80, 141-51.*

[11] Fitzmaurice, G.M., Laird, N.M. and Ware J.H. 2004.Applied longitudinal data. *J. Wliley & sons. pp 326-328.*

[12] Hall, B. and Severini, T. 1998.Extended Generalized Estimating Equations for clustered data. *J. Amer. Stat. Assoc. 93, 1365-1374.*
[13] Harville, D.A. 1977. Maximum likelihood approaches to variance components estimation and to related problem. J. Amer. Stat. Assoc. 72, 320-329.

[14] Hu, M. and Lachin, J.M. 2001.Application robust estimating equations to the analysis of quantitative longitudinal data. *Stat. in Med. 20, 3411-3428.*

[15] Jennrich, R.I. and Scluchter, M.D. 1986.Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics 42, 805-820.*

[16] Laird, N.M and Ware, J.H. 1982. Random effect models for longitudinal data. *Biometrics 38, 963-974.*

[17] Laird, N.M., Lange, N. and Stram, D. 1987. Maximum likelihood computations with repeated measures: application of the E-M algorithm. *J. Amer. Stat. Assoc. 82, 97-105.*

[18] Liang, K.Y. and Zeger, S.L. 1986.Longitudinal data analysis using GLM. *Biometrika 73, 13-22.*

[19] Liang, K.Y., Zeger, S. L. and Qaquish, B. 1992.Multivariate regression analysis for categorial data. *J. Roy. Stat. Soc., B 54, 3-40.*

[20] Lindsey, J.K. and Lambert, P. 1998.On the appropriateness of marginal models for Repeated Measurements in Clinical Trials. *Stat. in Med.17, 447-469.*

[21] Lindstrom, M.J. and Bates, D.M. 1988. Newton-Raphson and E-M algorithm for linear mixed effect models for repeated -measures data. *J. Amer. Stat. Assoc. 83, 1014-1022.*

[22] Littell, R.C., Milliken, G.A.,Stroup, W.W. and Wolfinger, R.D. 1996.*System for mixed models. SAS institute inc: Cary,N.C.*

[23] Littell, R.C., Pendergast, J. and Natarajan, R. 2000.Modelling covariance Structure in the analysis of repeated measures data. *Stat. in Med. 19, 1793-1819.*

[24] 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.*

[25] McCullagh, P. and Nelder, J.A. 1989.Generalized linear model.*2nd edition. London: Chapman and Hall.*

[26] Nelder, J.A. and Pregibon, D. 1987.An extended quasi-likelihood function. *Biometrika 74, 221-232.*

[27] Park, T. 1993. A comparaison of the GEE approach with the maximum likelihood approach for repeated measurements. *Stat. in Med. 12, 1723-1732.*
[28] Park, T. and Shin, D-Y. 1999. On the use of working correlation matrices in the GEE approach for longitudinal data. *Commun. in Stat.-Simul. 28, 1011-1029.*

[29] Prentice, R.L. 1988.Correlated binary regression with covariates specific to each binary observation. *Biometrics 44, 1033-1048.*

[30] 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.*

[31] Verbeke, G. and Molenberghs, G. 2000. Linear Mixed Models for longitudinal data. *New York: Springer*.

[32] Wedderburn, R.W.M. 1974.Quasi-likelihood function; GLM and the Gauss- Newton method. *Biometrika 61, 439-447.*

[33] Zhao, L.P. and Prentice, R.L. 1990.Correlated binary regression using a quadratic exponential model. *Biometrika 77, 642-648.*

[34] Zeger, S.L and Liang, K.Y. 1992.An overview of methods for the analysis of longitudinal data. *Stat. in Med. 11, 1825-1839.*

[35] Zerbe, G.O. 1979. Randomization Analysis of Growth Curves. *J. Amer. Stat. Assoc. 74, 215-221.*