COMPSTAT

Proceedings in
Computational Statistics

14th Symposium held
in Utrecht,
The Netherlands, 2000

Edited by
Jelke G. Bethlehem
and Peter G.M. van der Heijden

With 117 Figures
and 96 Tables

Physica-Verlag
A Springer-Verlag Company

Markov Chain Monte Carlo methods for handling missing covariates in longitudinal mixed models

Ernst Schuster

Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Liebigstrasse 27, 04103 Leipzig, Germany.

Abstract. Handling missing covariates in longitudinal mixed effect models is demonstrated on a medical example.

Keywords. Markov Chain Monte Carlo, Missing Covariates, Longitudinal Mixed Models

1 Introduction

Regression models for longitudinal data with fixed and random effects for the covariates are investigated. One aim of statistical analysis is the determination of those covariates which have a significant influence on the response or which lead to a good model fitting.

Values of the response variable may be *missing at random* and we assume that the parameters of the data model and the parameters of the missingness mechanism are *distinct*, e.g. for the response variable the missing-data mechanism is *ignorable* (Schafer (1997)). If all covariates are complete, the observed-data likelihood can be utilized. But if covariates are missing it is intractable and multiple imputation is one possibility. Neither this method nor the EM algorithm, should be examined here. In my opinion the best way to handle missing covariates is the use of Markov Chain Monte Carlo (MCMC) methods to estimate the model parameters. As a result the missing covariates are estimated together with the whole model from a sample of the posterior distribution using non-informative priors for these. The analysis with complete covariates and the MCMC method with missing covariates is demonstrated on a medical example.

2 Statistical modelling and medical example

The examined model is a mixed effect regression model for repeated measurements with missing response values:

$$y_i = X_i\beta + u_i + \varepsilon_i$$

where y_i is an n_i column vector of the response variable for case i, X_i is an n_i x p design matrix with the values of the p covariates, β is a p column vector of regression coefficients assumed to be fixed, u_i are random intercepts, which are assumed to be independently distributed across subjects with $u_i \sim N(0, \sigma^2)$, ε_i is the

within subject error with $\varepsilon_i \sim N(0, W)$, where W, for within, is a covariance matrix – n_i x n_i.

The aim of a prospective study was to evaluate the usefulness of clinical parameters, laboratory tests and immunogenetic markers as predictors of an erosive course of joint disease early in rheumatoid arthritis (RA). 96 patients with persistent oligio-or polyarthritis over a period of six weeks and a history of disease of less than 2 years were enrolled in the study. As one major outcome parameter of the destructive process, radiological evaluation of joint erosions was used. At study entry and at each scheduled visit, hand and feet radiographs were taken and scored using the Larsen score. As response for statistic modelling, **changes** in the Larsen score in the first (96), second (96), third (72) and fourth (54) year of observation were used. In parentheses the number of measured changes in the Larsen score is specified. Because the study is going on, the missing values of the response variable can be treated as ignorable.

Since the goal of the analysis was the identification of prognostic parameters, initial values of the covariates documented at study entry as well as the values obtained after 6 months of observation and therapy were used. In the preselected model (Table 1) only rheumatoid factor IgA (RfIgA) has two missing values. These missing values are also handled as ignorable.

3 Markov Chain Monte Carlo methods

MCMC methods are generic simulation methods. The unknown parameters and the missing values get prior distributions, mostly noninformative e.g. $N(0, 10^6)$ for location parameters and $Gamma(10^{-3}, 10^{-3})$ for each precision, which is the reciprocal of variance. Using Bayes' theorem the posterior density is proportional to prior times likelihood. In case of noninformative prior densities the posterior density results in similar estimations as likelihood function does. Therefore a sample from the posterior density is used, when the observed-data likelihood is intractable. The most important MCMC method is the algorithm of Metropolis and Hastings (Robert, C.P., Casella, G. (1999)) with the Gibbs sampler (Gilks, W.R., and Roberts, G.O. (1996)) as special case. The Gibbs sampler can be used if it is possible to sample from the conditional densities of each parameter given all other parameters and the data.

By using WinBUGS the joint posterior must be conjugate or log-concave. Otherwise, they are discretized. The Gibbs Sampler, after reaching stationary distribution, can be used for inference.

4 Modelling and results for the medical example

4.1 Complete covariates analysis and selection of the covariance structure

Mixed effect regression models were calculated with S-Plus 2000 for all cases with complete covariates. The stepwise term reduction began for a model with the initial values documented at study entry as well as the values obtained after 6 months of observation and therapy as covariates for fixed effects.

Table 1. Parameter-estimator (with p-value in parenthesis) for the pre-selected model, models with other covariance-structure and the BUGS-model

		Pre-selected model	Conditional independent + random slope	Conditional independent model	WinBUGS model
		estimator (p-value)	estimator (p-value)	estimator (p-value)	Estimator: median (95% credible interval)
Fixed effects	intercept	12.46 (<.0001)	13.34 (<.0001)	13.32 (<.0001)	13.31 (8.53, 18.27)
	DR4epi	2.87 (0.0087)	2.74 (0.0069)	2.77 (0.0059)	2.80 (0.84, 4.73)
	tint	-2.92 (0.0006)	-3.31 (0.0004)	-3.30 (0.0004)	-3.28 (-5.10, -1.56)
	Sex	-7.69 (0.0014)	-9.11 (0.0019)	-9.05 (0.0012)	-8.86 (-14.21,-3.72)
	RflgA	0.0095 (0.0162)	0.0114 (0.0021)	0.0107 (0.0032)	0.0094 (0.0026,0.016)
	Sex:tint	2.11 (0.0245)	2.72 (0.0075)	2.70 (0.0080)	2.65 (0.75, 4.67)
Random effects: standard deviation	intercept	3.06	4.21	2.29	2.32 (0.35, 3.70)
	slope		0.74		
	Residual	7.08	6.97	7.10	7.18 (6.55, 7.90)
Correlations between time intervals	r_{12}	-0.237			
	r_{13}	0.127			
	r_{14}	0.158			
	r_{23}	0.007			
	r_{24}	0.102			
	r_{34}	0.556			
LR-Test to cond. ind. model: p-value		0.0406	0.3434	---	
AIC		2165	2168	2166	
BIC		2218	2206	2196	

Some interactions were also included. The stepwise deletion of parameters (biggest p-value first), that had no effect on the yearly increase of Larsen score leads to a preselected model. In this example the Akaike Information Criterion (AIC: -2 log-likelihood + p * 2) and the Bayesian Information Criterion (BIC: -2 log-likelihood + p * log(number of cases)) are minimized, too. Table 1 shows the preselected model in the first column. Conditional on the selected covariates and the random intercept the correlations between time intervals are nearby zero (exception r_{34}). Therefore it makes sense to use a conditional independent model (Table 1, third column). This is equivalent to the compound symmetry form for the total

covariance matrix. The conditional independent model is worse for the log-ratio-test (p=0.04) but better in BIC. The fit of both models is equally good, therefore the more simple conditional independent model should be preferred, which is analysed with WinBUGS below. The second column shows that an additional random slope is not necessary, because AIC and BIC are worse.

4.2 Modelling with MCMC methods for missing covariates

Fig 1 shows a directed graph for the model created with WinBUGS. The graph is a special case of the general relation

$$\mathbf{y}_i \sim N(\mathbf{X}_i \beta + u_i , \sigma^2 I)$$

based on the following distributions and equations:

$$\mathbf{y}_i \equiv (larsd_{i1}, larsd_{i2}, larsd_{i3}, larsd_{i4})^T,$$

$$\mathbf{X}_i \beta \equiv \alpha_0 + b_1 tint_k + b_2 Sex_i + b_3 Sex_i tint_k + b_4 DR4epi_i + b_5 RfIgA_i ,$$

$$\beta \equiv (\alpha_0 , b_1 , b_2 , b_3 , b_4 , b_5)^T , \qquad \mathbf{tint} = (1, 2, 3, 4)^T ,$$

$$u_i \equiv \sigma_1 \alpha_i , \qquad \mu_{i,k} \equiv \mathbf{X}_i \beta + u_i ,$$

$$larsd_{ik} \sim Normal(\mu_{i,k} , \tau), \qquad larsd_{ik} \sim Normal(0, 10^{-6}) \text{ for missing } larsd_{ik}$$

$$\alpha_0 , b_1 , ..., b_5 \sim Normal(0, 10^{-6}) , \tau \sim Gamma(10^{-3} , 10^{-3}) ,$$

$$\sigma_1 \sim Unif(0, 10) , \qquad \alpha_i \sim Normal(0, 1) ,$$

$$RfIgA_{11} \sim Unif(0, 1000) , \qquad RfIgA_{45} \sim Unif(0, 1000) .$$

In contrast to the notation $N(\mu, \sigma^2)$, $Normal(\mu, \tau)$ means here and in BUGS normal distributed with expectation μ and precision τ, which is the reciprocal of the variance. The fixed effects are the constant α_0 , the slope in time b_1, and the other factors b_2 to b_5 for the covariates. All these get as prior a noninformative normal distribution.

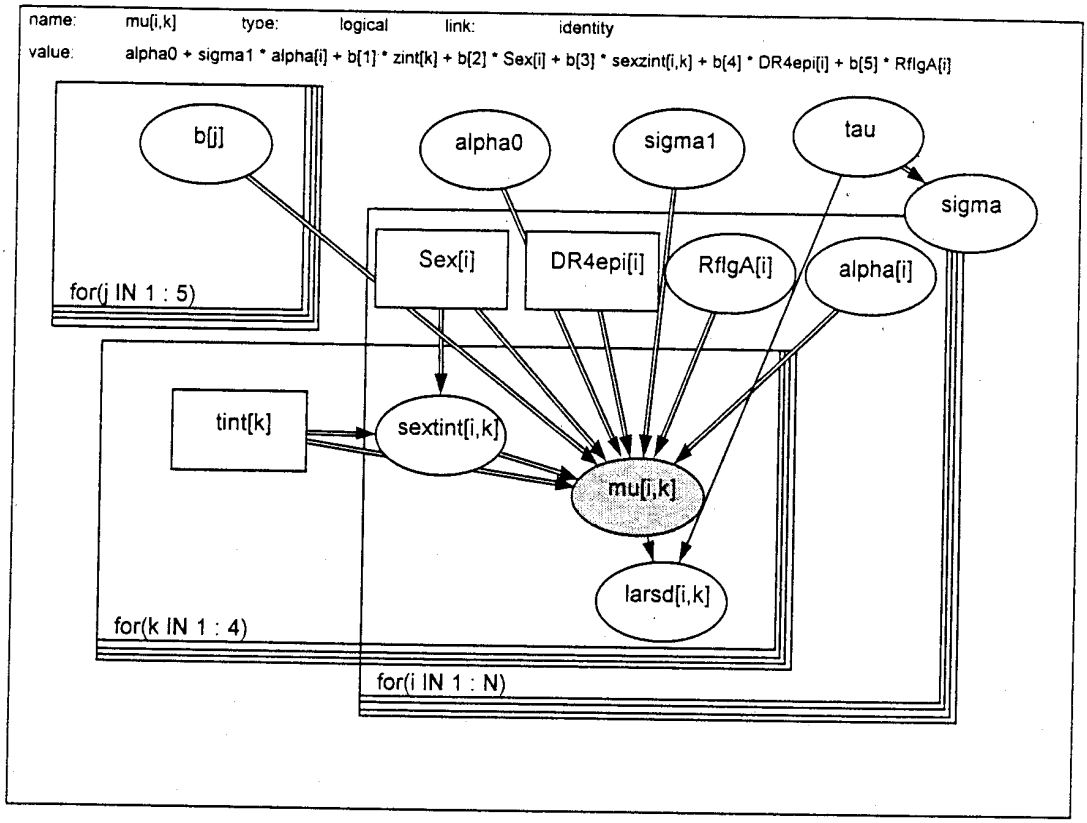

Fig 1. Graphical model of the rheumatoid arthritis data

The factor σ_1 is the standard deviation of the random intercept. Therefore the α_i are standard normal distributed. The factor σ_1 and both missing values of RfIgA get as prior a noninformative uniform distribution. The yearly change in Larsen score (larsd) is modelled as normal distributed. As prior for τ is chosen a noninformative conjugate distribution, namely a gamma distribution.

Table 1 shows in the last column the result of 50 000 iterations after a "burn in" of 1 million iterations. Because only two values of RfIgA were missing, the conditional independent model with complete covariates and the BUGS model are rather similar.

In the BUGS model, the progression of disease is indicated by a yearly baseline increase in the Larsen score of 13.31, that is described by the intercept. The influence of time indicates that this increase decreases by 3.28 yearly. The presence of an RA associated DRB1*04 allele had the strongest impact on progression resulting in an additional yearly increase in Larsen score of 2.8. In addition, the level of RFIgA measured at study entry had a significant influence on the yearly increase in Larsen score. There was a gender difference with men having a higher yearly increase of 8.86 compared to women. However the interaction with time interval shows that this increase decreases by 2.65 yearly.

A remarkable benefit of the MCMC methods is the ability to calculate credible intervals for random coefficients. A crucial advantage of MCMC methods is the possibility to calculate robust models by other distributions e.g. double exponential for the error term.

The results show that for data with missing covariates the use of MCMC methods can be recommended.

Acknowledgements. I am grateful to Dr Sylke Kaltenhäuser of the Department of Medicine IV, University Leipzig for permission to use the rheumatoid arthritis data.

References

Brooks, S.P., (1998). Markov Chain Monte Carlo method and its application. *The Statistician*, **47**, 69-100.

Gamerman, D. (1997). *Markov Chain Monte Carlo – Stochastic simulation for Bayesian inference*, London: Chapman and Hall.

Gilks, W.R., and Roberts, G.O. (1996). Improving MCMC mixing. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter). London: Chapman and Hall.

O'Hagen, A. (1994). *Kendall's Advanced Theory of Statistics, Vol 2B, Bayesian Inference,* London: Edward Arnold.

Robert, C.P., Casella, G. (1999) *Monte Carlo Statistical Methods*, New York: Springer.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schuster, E. (1999). Markov Chain Monte Carlo Methods for Handling Missing Covariates in Multiple Regression Models. In: Friedl, H. Berghold, A. and Kauermann, G. (eds.) *Statistical Modelling – Proceedings of the 14th International Workshop on Statistical Modelling, Graz, Austria, July 19-23, 1999*, 651-655.

Spiegelhalter DJ, Thomas A, Best NG. (1999). *WinBUGS Version 1.2 UserManual. Cambridge, U.K.*

S-PLUS 2000 Guide to Statistics *(1999), MathSoft, Seattle.*

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Third Edition. New-York: Springer.