Statistical Modelling

Proceedings of the

14th International Workshop on Statistical Modelling

Graz, Austria, July 19-23, 1999

H. Friedl, A. Berghold, G. Kauermann
(Editors)

Markov Chain Monte Carlo (MCMC) Methods for Handling Missing Covariates in Multiple Regression Models

Ernst Schuster[1]

[1] Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Liebigstr. 27, 04103 Leipzig, Germany; Schuster@imise.uni-leipzig.de

Abstract: In the classical multiple regression model one proceeds on the assumption that the data are complete. In practical applications however, because of the restriction to complete cases, a situation may occur where only less than half of all cases can be considered. This may lead to a considerable bias in the results. One possible solution has been offered by the Expectation-Maximization algorithm, known as EM algorithm (Dempster et al., 1977), however, this method may lead to local maxima of the likelihood function and therefore I cannot recommend it. At present Markov Chain Monte Carlo (MCMC) methods seem most suitable for consideration of missing covariates. The application of the suggested MCMC methods is demonstrated by using the WinBUGS software (Spiegelhalter et al., 1998) on a medical data set.

Keywords: Markov Chain Monte Carlo; MCMC, Missing Covariates; Multiple Regression Models.

1 Introduction

One aim of multiple regression analysis is the determination of those covariates which have a significant influence on the response or which lead to a good model fitting by stepwise methods (forward, backward, or both). As a rule only cases with complete set of covariates given can be used for computations in multiple regression models. Therefore a considerable part of information available may not be included if any covariate is missing. Some software (e.g. SPSS) offer the possibility to compute the correlation coefficients used in regression analysis from all pairs of values available. So the cases with missing covariates will be included into the analysis at least partially. If these computations, however, result in a very different model (as in the example below) a more detailed handling of missing data should be performed. For such a problem the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) is suitable, however, this method may lead to local maxima of the likelihood function and therefore I cannot recommend it. Alternatively MCMC methods can be used which are based on Bayesian statistics. In this method the unknown parameters and the missing values get non-informative a priori distributions, namely 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 determined by

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\theta)p(x|\theta)d\theta}$$

The posterior density is proportional to prior times likelihood. In case of non-informative prior densities the posterior density results in similar estimations as likelihood function does. The computation of the normalizing constant above is rather difficult because of the necessity of high-dimensional numeric integration. This can be avoided by using the MCMC methods in order to sample without knowing the normalizing constant. In this case the missing values are treated as additional parameters which get non-informative prior density and they are estimated, too. By applying this method cases with missing covariates can also be included. The WinBUGS software (Spiegelhalter et al., 1998) was used for performing MCMC methods.

2 Markov Chain Monte Carlo Methods

The most important MCMC method is the algorithm of Metropolis and Hastings (Metropolis et al., 1953) with the Gibbs sampler (Gelfand and Smith, 1990) 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. These full conditional densities will be explained in the following presentation. By using WinBUGS these densities need not be specified explicitly because WinBUGS performs the derivation. The Gibbs sampler cyclically draws random values from these one-dimensional conditional distributions which as a rule are known conjugated distributions from which samples can be drawn. The Gibbs Sampler generates random points $\theta_{1,}$ $\theta_{2,}$ ... (1st index (before comma) index of Markov chain, 2nd index (after comma) component of vector θ). The starting point $\theta_{0,} = (\theta_{0,1}, \ldots, \theta_{0,p})$ should be given arbitrarily. $\theta_{m+1,}$ results from $\theta_{m,}$ by the following formulas

$$\begin{array}{lll}
\theta_{m+1,1} & \text{is drawn randomly from} & f_1(\theta_1|\theta_{m,2}, \ldots, \theta_{m,p}) \\
\theta_{m+1,2} & \text{is drawn randomly from} & f_2(\theta_2|\theta_{m+1,1}, \theta_{m,3}, \ldots, \theta_{m,p}) \\
\vdots & \vdots & \vdots \\
\theta_{m+1,p} & \text{is drawn randomly from} & f_p(\theta_p|\theta_{m+1,1}, \ldots, \theta_{m+1,p-1})
\end{array}$$

The Gibbs sampler converges (O'Hagen, 1996) asymptotically under regularity conditions, if the chain is irreducible and aperiodic, against the posterior density which was searched for. Therefore after reaching stationary distribution, the Gibbs Sampler results in an independent sample of the normalized joint distribution $f(\theta)$; i.e. of the distribution of the parameters given the data. This sample can be used as any random sample. The sample means of the parameters are estimators of the corresponding expectations. The variance of the parameters can be estimated from the same sample. The same is valid for any function $h(\theta)$. From the sample one can get in addition the marginal densities, for example with kernel estimators.

3 Example: Prognostic Factors Influencing the Clinical Course of Recent Onset Rheumatoid Arthritis

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 the course of rheumatoid arthritis (RA). 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. Morning stiffness, number of swollen joints and joint tenderness score (Ritchie index: RI), standard lab tests (ESR, CRP, fibrinogen and quantitative determination of IgM and IgA rheumatoid factors: RfIgM and RfIgA) and flowcytometric markers of lymphocyte activation including the ratio of CD4/CD8 positive T cells (CD4.8) were documented 6-monthly initially, and every year later in the study. HLA-DRB1-alleles and DR4 subtypes (DR4epi) were determined by hybridisation of PCR products to sequence-specific oligonucleotypes. Hand and feet X-rays were taken yearly and judged according to Larsen's method. Larsen Indices scored by the patients at study entry and subsequently after one and two years of observation were available for 88 patients, while 48 patients were followed for four years.

As response for statistic modelling, changes in the Larsen score in the first and second year and the averaged yearly change during the third and fourth year of observation (y) were used. Since the goal of the analysis was the identification of prognostic parameters, initial values of the covariates documented at study entry (denoted with 1 at the end of the variable name) as well as the values obtained after 6 months of observation and therapy (denoted with 2 at the end of the variable name) were used. The appropriated model for the data presented here is, therefore, a mixed effect regression model for repeated measurements. After initial analysis, a model with random intercept and compound symmetry was chosen to describe the course of the disease. For statistical analysis, the S-Plus 4.5 software with the functions lme and lm was used. Since serial correlation between different time points was rather low ($r = -0.04$), separate observations could also be treated as independent, which leads to a multiple linear regression. Therefore these data can serve as an example here. The backward and forward stepwise solution of this regression results in an exactly identical model (REG1)(with 95% confidence intervals for the parameters in parentheses):

$$\begin{aligned} y \;&=\; 3.73(1.20,\ 6.25)^{*}\ \text{DR4epi} - 3.17(-5.91,\ -0.43)^{*}\ \text{Time01}\\ &\quad + 0.041\ (0.013,\ 0.069) * \text{RflgA2} + 2.15(-0.17,\ 4.46) * \text{CD4.8.1}\\ &\quad - 3.83(-6.50,\ -1.16) * \text{CD4.8.2} + 7.13(3.46,\ 10.81) \end{aligned}$$

Preceding investigations using the time as factor showed that it is sufficient to distinguish between the first two years (Time01=0) and the following two years (Time01=1). Regression analysis with computation of correlation coefficients in pairs results in (REG2):

$$\begin{aligned} y \;&=\; 3.29(0.78,\ 5.81) * \text{DR4epi} - 2.89(-5.45,\ -0.34) * \text{Time01}\\ &\quad - 3.00(-6.31,\ 0.31) * \text{Sex} + 0.014(0.004,\ 0.024) * \text{RflgA1}\\ &\quad - 2.00(-3.82,\ -0.18) * \text{CD4.8.2} + 10.69(6.27,\ 15.11) \end{aligned}$$

Male patients are coded as Sex=0 and female ones as Sex=1. An analysis without covariates CD4.8 containing more then 20 missing values, results for both kinds of computation of correlation coefficients in (REG3):

$$\begin{aligned} y \;&=\; 3.14(0.87,\ 5.42) * \text{DR4epi} - 3.08(-5.73,\ -0.43) * \text{Time01}\\ &\quad - 2.635(-5.517,\ 0.247) * \text{Sex} + 0.012(0.003,\ 0.022) * \text{RflgA1}\\ &\quad + 0.202(-0.015,\ 0.418) * \text{RI2} + 6.02(2.84,\ 9.20) \end{aligned}$$

By using WinBUGS the missing covariates get a non-informative prior density and are also estimated, with the effect that all 223 time points can be included into the analysis. After the input of the model and the data WinBUGS independently selects the update methods. Subsequently the initial values for the parameters must be specified by the user or by WinBUGS. After a "burn in" of at least 10 000 iterations the Markov chain may considered as converging in all computed examples. This was verified by the examination of the iteration sequence and by using the program CODA (see Best et al., 1995). From the drawn sample of size 10 000 the medians and the 95% credible intervals were estimated. The results show that a reduction of variables is necessary which leads to the model BUGS1 step by step:

$$\begin{aligned} y \;&=\; 3.29(0.95,\ 5.50) * \text{DR4epi} - 2.98(-5.57,\ -0.38) * \text{Time01}\\ &\quad - 2.92(-5.77,\ -0.07) * \text{Sex} + 0.012(0.003,\ 0.021) * \text{RflgA1}\\ &\quad + 7.24(4.35,\ 10.09) \end{aligned}$$

In BUGS1 the response y, the yearly increase in the Larsen score, was modelled as normally distributed with mean mu and precision tau. The use of a normal distribution for y corresponds to an approximately squared adjustment of the residuals. Because of the large deviations in the residuals of y it would be better to use an adjustment in absolute values. Maximum likelihood estimates for the double exponential distribution are essentially equivalent to minimising the sum of absolute deviations, see Birkes and Dodge (1993). Therefore the double exponential distribution with the density $(\tau/2)\,\mathrm{e}^{-\tau|y-\mu|}$ is also used for y. The reduction of variables leads to the model BUGS2,

y = 3.55(2.04, 5.31) * DR4epi - 2.33(-4.22, -0.69) * Sex
+ 0.014(0.008, 0.023) * RfIgA1 + 2.48(0.86, 4.34)

which differs from BUGS1 only in the fact, that the time is not in the model. It seems that BUGS2 is fitted best to the data and therefore this model is the model of choice. The presence of a RA associated DR4 allele resulted in an additional averaged yearly increase of 3.6 points in the Larsen score. Men have a by 2.3 points higher yearly increase in the Larsen score as compared to women. In addition, RF-IgA measured at study entry had a significant impact of yearly increase on the Larsen score. It is concluded that the developed model describes the course of the early phase of RA. It can reliably be used for the first 5 years. For the later course no data are available, however, linear models seem not suitable because the Larsen score is limited. Therefore saturation curves should be taken into consideration. This paper demonstrates the superiority of MCMC methods in modelling complex data structures with missing values.

Acknowledgements: I am grateful to Dr Sylke Kaltenhäuser of the Department of Medicine IV, University of Leipzig for permission to use the Rheumatoid Arthritis Data.

References

Best, N.G., Cowles, M.K., Vines, S.K. (1995). *CODA Manual version 0.30.* MRC Biostatistics Unit, Cambridge, UK.

Birks, D., Dodge, Y. (1993). *Alternative Methods of Regression.* John Wiley and Sons, New York.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete date via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B*, **39**, 1-38.

Gelfand, A.E., Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

McLachran, G.J., Krishan, T. (1997). *The EM algorithm and extensions.* New York: John Wiley & Sons.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

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

Spiegelhalter, D., Thomas, A., Best, N. (1998). *WinBUGS user manual Version 1.1.1*

S-PLUS 4 Guide to Statistics (1997), MathSoft, Seattle.

SPSS Base 8.0 for Windows User's Guide (1998), SPSS Inc.