Award no. R000222732

Award Holders: Prof. H. Goldstein, J. Rasbash

Title: Developing graphical and inferential tools for social data analysis

Full Report on Research Activities and Results

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Full Report of Research Activities and Results

Background

In recent years, due partly to the increasing role of the computer in society, methodology and software have become available which allow researchers to study statistical models with increasingly complex data structures. The ESRC ALCD programme has helped with the development of both MCMC and likelihood based methods for fitting multilevel models to these complex data structures, in part through their support of the computer software packages BUGS (Spiegelhalter et a!. 1997) and MLwiN (Rasbash et al. 1999).

Both of these software packages have increased the group of researchers that can fit complex statistical models from a small group of statistical methodologists to the social and medical science research community in general. This increase in practitioners of multilevel analyses has thrown up models of increasingly higher levels of complexity, and has lead to new insights in a number of areas such as education, geography, demography and political science. The increase in interest is leading to models of greater complexity and a myriad of different statistical estimation methods to fit these new models.

Examples of statistical models that extend the basic multilevel models, include cross classified models, where for example, in education, pupils are classified by both their primary and secondary school attended but primary school is not nested within secondary school (Goldstein and Sammons, 1997). This can be extended to include multiple membership models where pupils may move from school to school and so the effect of the several, say secondary, schools attended on a pupil's response variable must be accounted for.

Statistical methodologists now have a vital, multifaceted role in the progression of the field of multilevel modelling. They firstly need to devise new methods, or adapt existing methods to fit the new complex data structures. They then need to implement their methods in statistical software so that the methods can be used by the social science research community. Comparison work comparing the various approaches to fit these models is also required, so that they can advise the user community on which methods to use and how and when to use them.

The first version of the software package MLwiN released to the user community in February 1998 had many advantages over its predecessor MLn. It was for the first time a Windows based multilevel modelling package that included better graphical functionality and an equation based interface. It was also the first package that included both maximum likelihood and Bayesian MCMC estimation methods for fitting multilevel models (Browne 1998), and also included another simulation-based technique, bootstrapping (Efron and Tibshirani 1993) to fit multilevel models in MLwiN.

Although all these methods exist, there is very little in the statistics literature to compare these various methods when fitting multilevel models. This project aimed to both bring these simulation based methods to the attention of the social science community and to provide some advice on when to use each method.

Objectives

This research project has had 3 main aims which have been persued as follows:

1. **To provide guidance to users on the use of these techniques using the graphical user interface in MLwiN together with any necessary enhancements.**

In the first release of MLwiN, the MCMC and bootstrap estimation methods were introduced together with two new graphical windows. The trajectories window allows the user to follow the cycles of the MCMC or Bootstrap iterations to obtain visual feedback on performance, and in particular to detect quickly where major estimation difficulties may be occurring. The diagnostics window, accessed from the trajectories window, presents the user with detailed information about each parameter being estimated. It provides Markov chain 'traces' and estimates of quantiles and chain lengths for satisfactory convergence. The project has carried out simulation studies to compare the various estimation methods available for fitting multilevel models, concentrating on comparing the likelihood-based methods and the various Bayesian MCMC methods for a few of the most common multilevel models. The procedures used in setting up the simulation studies will be outlined in the Methods section of this report. Following the simulation studies and discussion with the user community through both workshops and meetings, modifications and enhancements have been made to many of the features of the graphical interface, and these are being incorporated into the second major release of *MLwiN* in late 1999. These modifications will be summarised in the Results section of this report and the results of the simulations are described in two forthcoming joint papers, Browne and Draper
(1999A, 1999B).

2. **To provide a body of exemplar applications to guide users.**

The manual that accompanied the first release of MLwiN contained only two short chapters on the use of the simulation-based methods. During the course of this project an updated version of the user manual has been written to complement the second major release of the software. This new manual both introduces the new features in the software and expands on the existing material on the simulation methods. The new manual has three sections, with the third section solely concerned with the MCMC and bootstrap methods. This section contains an introductory chapter on simulation-based methods to motivate new users in their use. There are then 3 chapters on the MCMC methods which include more of the theory surrounding the methodology as well as greater explanation of the various MCMC methods and options in the software package. The bootstrapping chapter has also been expanded to include a new non-parametric bootstrapping method based on sampling from modified residuals.

3. **To disseminate knowledge about these techniques through workshops, conferences and via the internet.**

During the course of the project Dr Browne has assisted on 6 MLwiN workshops which have incorporated the simulation based methods, including 1 workshop based solely on the MCMC options in the package. This provided an opportunity to show new users that there are alternative methods, both classical and Bayesian, for fitting multilevel models and to explain to them a little of the theory and motivation behind these methods. These workshops were not only a good means of dissemination for the project but also were good opportunities for getting opinions from the user community.
During the course of the year Dr Browne and Professor David Draper, wh& acted as a consultant to the project, gave talks at the 2nd International Amsterdam conference on Multilevel Analysis

based on two forthcoming joint papers that were well received. Talks were also given to both the health economics group at the University of York and the MRC social and public health sciences unit at the University of Glasgow on MCMC estimation methods for fitting multilevel models.

Before the project was started there was an existing web site for the MLwiN package (http://www.ioe.ac.uklmlwin) which contained details of the software and was used to inform users of developments as well as allowing them to download upgrades of the software. Dr Browne took over responsibility for the 'bug' reporting pages on this web-site which report known problems in the package and in which version they are fixed. During the pre-Beta testing period of the second release of MLwiN (version 1.1) a web site has been created to enable the researchers who are testing the software to download the latest versions of the software, manual chapters and help system. This web site also includes a discussion list to allow testers to post problems and for us to announce new improvements. It is anticipated that this web site will be a template for a full Beta release web site in October/November 1999.

Methods

To achieve the first aim of providing users with advice on the appropriate method to use for their individual multilevel dataset three large simulation studies were set up. The first simulation study considered the simple 2 level variance components model to compare the IGLS and RIGLS maximum likelihood methods with MCMC methods using different 'non-informative' prior distributions. The study was designed around the JSP dataset (Mortimore et a!. 1988) and several factors were tested. Firstly the effect of study design was tested by altering the number of level 1 and 2 units and whether the study was balanced or unbalanced. Secondly the underlying true values of the level 1 and 2 variance to be simulated from were modified. In all this gave 15 simulation scenarios and for each scenario 1,000 datasets were generated and fitted using the various methods.

The second simulation study considered logistic regression multilevel models. This study considered some simulation datasets created in Rodriguez and Goldman (1995) that had highlighted deficiencies in quasi-likelihood estimation methods. These datasets were used to compare the quasi-likelihood methods MQL and PQL with the MCMC methods considered in the first simulation study. Both of these simulation studies are covered in more detail in Browne (1998) and Browne and Draper (1 999A).

The third simulation study that is again based on the JSP dataset was set up to consider random slopes regression models. This study was designed to consider the choice of 'non-informative' priors for variance matrices. In all 9 study designs were chosen to compare different sizes of study, balanced and unbalanced designs and different correlations in the level 2 variance matrix. As with the variance components model 1,000 datasets were generated for each simulation design. More details on this study, along with some comparison work between different MCMC samplers can be found in Browne (1998) and Browne and Draper (1999 B). In these studies parametric bootstrap methods for non-Bayesian models were not included since their properties have already been explored and the results incorporated into the existing version *of MLwiN* (Goldstein et al., 1998). They provide small sample inferences with desirable properties, but their major drawback is that they tend to be computationally intensive.

Results

The general conclusions that have come out of the simulation studies are that there are uses for both maximum likelihood based methods and Bayesian methods. The likelihood based methods are far quicker than the Bayesian methods but the studies show that the Bayesian methods can match or outperform likelihood methods in terms of both estimate bias and interval coverage.

Following the simulation studies and discussions with the user community the following changes, reflecting these results, have been made to the MCMC options in the MLwiN software package.

- There is a large variability in the knowledge of MCMC methods amongst the users of MLwiN. Consequently the MCMC interface now has two forms to accommodate the two types of user.
- For the new user, the Estimation control window from which the user chooses between the various estimation methods now has a general MCMC method. For this method the software will choose the appropriate default values for each of the MCMC specific options and the MCMC estimation methods used for each group of parameters based on the type of model fitted. This means that new users can use the MCMC methods without being overwhelmed with new terminology and lots of options.
- The simulation studies carried out in this project (Browne (1998) and Browne and Draper (1999A, 1999B)) suggest that it is best to use Gibbs sampling for all parameters in a Normal response model. For discrete response models the adaptive Metropolis Hastings (MH) method performs favourably with the adaptive rejection methods used in BUGS (Spiegelhalter et al. 1997) for both the residuals and fixed effects (using Gibbs sampling for variance parameters). These will be the default estimation methods mentioned above.
- For the advanced MCMC user there is an Advanced MCMC options screen that will allow the user to select their own preferred estimation method for each group of parameters (where a choice is offered in MLwiN). They will also be allowed to change the settings associated with the Metropolis Hastings method and to experiment with the new multivariate normal proposal MH sampler for some parameters. All these new features are described in the user guide.
- The simulation studies also suggested using different default ‘uninformative’ prior distributions for the variance parameters. Consequently in the new version of MLwiN the default ‘uninformative’ prior for the variance parameters has been changed to an inverse gamma distribution. The default uniform priors on the variance scale, used in the first version of *MLwiN* , are available as an alternative and can be chosen from the advanced MCMC options screen.
- The MCMC diagnostics screen has been updated to include some additional convergence diagnostics. The Raftery-Lewis diagnostic (Raftery and Lewis 1992) is now calculated at the two end points of an interval rather than at a single point. This allows the user to check the behaviour of the chain in both tails of the distribution. The Brooks-Draper diagnostic (Brooks and Draper 1999) has also been included. This is a diagnostic that gives estimated run lengths based on the mean of the chain rather than the quantiles. An additional MCSE (Monte Carlo standard error) graph completes the new features and gives the user an indication of how long they will need to run their model to obtain estimates with a given MCSE.
- The default settings for the plots and diagnostics found on the MCMC diagnostics screen can now be altered via an additional options screen. This gives the user greater flexibility in analysing their parameter chains.

Activities

The MLwiN project team organise 'fellows meetings' every month, in which the core team members are joined by other researchers who have worked and continue to work closely with team through the ESRC ALCtTvisiting fellows scheme. These meetings have provided an opportunity for Dr Browne to present his work and get feedback and opinions from an audience of experienced MLwiN users. There are also introductory workshops occurring roughly every 3 months and each of these workshops is attended by around 25 new users of the MLwiN package. These workshops allow the introduction of simulation-based techniques to inexperienced users and establish what this user group thinks of the current software. At the 2nd International Amsterdam multilevel modelling conference in March 1999 Dr Browne talked about MCMC methods in general and this meeting provided an opportunity to converse with other experienced multilevel modellers. Seminars were also given at both York and Glasgow universities to mixed audiences that have included both MLWiN users and Bayesian statisticians. These seminars have introduced Bayesian MCMC methods to users of MLwiN as well as publicising the work of the project amongst the Bayesian statistics community.

Outputs

There are three main outputs from the work on this project.

- **The MLwiN software package**

The new release of the software package contains many new MCMC features as detailed in the Results section of this report. It is hoped that these new features will make the simulation-based methods easier to use and encourage more users to try the simulation-based methods. The MLwiN package is now used by over 1,500 researchers and research groups world-wide and is in use in postgraduate courses. It is hoped that with the new release this number will grow. It is also hoped that through the other publications detailed below, and the dissemination through workshops and seminars, more users will be encouraged to try the simulation-based methods.

- **The MLwiN user guide (version** 2.0)

The new version of the MLwiN user guide is far more comprehensive than the original and contains many improvements based on feedback from the introductory workshops where it is used. The simulation-based methods now have their own section of 5 chapters. This is far longer than in the original, not only to cover additional features, but to also include more theory and examples to make the features more user-friendly.

- **Two journal** articles, Browne and Draper (1999A) and Browne **and Draper (1999B).**

These articles present some of the first comparative work between maximum likelihood based methods and Bayesian methods in the field of multilevel modelling. They also include some comparisons between different MCMC approaches to fitting multilevel modelling.

- **Under- and over- dispersed models.**

A start has been made in investigating the incorporation of parameters describing under- and over- dispersed data in discrete response models. In collaboration with Dr David Clayton the project has developed an MCMC/Data Augmentation procedure for fitting such models for the Probit link ftrnction. This work is being prepared for publication and extensions are being studied.

Impacts

The major impact of the changes to the user interface of MLWiN will be seen later in the year when the new release of the software occurs. At present the MLwiN software package is widely used and there are already many journal articles published that quote the software. The feedback from the introductory workshops has been generally favourable with many people considering using the simulation-based methods on their own datasets.

Future Research Priorities

The comparison work between estimation methods started in this project will be extended in a new 3-year ESRC grant (R0002381 17) that will investigate more complex multilevel data structures. Some preliminary work in this area has been carried out in the current project and is detailed in Goldstein et al. (1999).

Through the talks at both Glasgow and York, it is anticipated that Dr Browne will collaborate with researchers at these two establishments on fitting more complex models using MCMC modelling. It is hoped that there will be collaboration with Dr Alastair Leyland on fitting multilevel spatial models using MCMC methods, following on work done using maximum likelihood methods in Leyland et. al (1999) and Langford et al. (1999). It is also hoped that there will be collaboration with Dr Nigel Rice on fitting a multilevel model to a health and wages multivariate response with selective dropout following on from work done by Chib and Hamilton (1999).

References

Brooks SP and Draper D (1999) Comparing the efficiency of MCMC samplers. Technical report, Department of Mathematical Sciences, University of Bath, UK.

Browne WJ (1998) *Applying MCMC Methods to Multilevel Models*, PhD dissertation, Department of Mathematical Sciences, University of Bath, UK.

Browne WJ and Draper D (1999A) A comparison of Bayesian and likelihood methods for fitting multilevel models. (Submitted to the Journal of the Royal Statistical Society Series B)

Browne WJ and Draper D (1999B) Implementation issues in the Bayesian fitting of multilevel models. *Computational Statistics*, under review.

Chib S and Hamilton BH (1999) Bayesian Analysis of Cross-Section and Clustered data selection models. Mimeo, Olin School of Business, Washington University.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W.S., Yang, M., Woodhouse, G., and Healy, M. J. R. (1998). *A user's guide to MLwiN*. London, Institute of Education:

Goldstein H and Sammons P (1997) The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. School effectiveness and school improvement. 8: 2 19-230.

Goldstein H, Woodhouse G, Browne WJ and Rasbash J (1999) Multilevel models in the study of population structures. *Presented at the European Population Conference,* The Hague.

Langford IH, Leyland AH, Rasbash J, Goldstein H, Day RJ and McDonald A (1999) Multilevel Modelling of Area-Based Health Data. In *Disease Mapping and Risk Assessment for Public Health.* Edited by A.B.Lawson *et al* pp 217-228 John Wiley & Sons.

Leyland AH, Langford IH, Rasbash J and Goldstein H (1999) Multivariate spatial models for event data. *Statistics in medicine, (to appear).*

Mortimore P, Sammons P, Stoll L, Lewis D, Ecob R (1988) *School Matters.* Wells: Open Books.

Raftery AE and Lewis SM (1992) How many iterations in the Gibbs Sampler? In J.M.Bernardo, J.O.Berger, A.P.Dawid, and A.F.M. Smith (Eds.), *Bayesian Statistics* 4, pp 763-773. Oxford: Oxford University Press.

Rasbash J, Browne WJ, Goldstein H, Yang M, Plewis I, Draper D, Healy M and Woodhouse G (1999). *A User 's Guide to MLwiN, Version 2.0,* London: Institute of Education, University of London.

Rodriguez 0 and Goldman N *(1995).* An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 158,* 73-89.

Spiegeihalter DJ, Thomas A, Best NG and Gilks WR (1997) *BUGS. Bayesian Inference Using Gibbs Sampling, Version 0.60.* Cambridge: Medical Research Council Biostatistics Unit.