International Journal of Science and Technology Education Research Vol. 1(2), pp. 32 - 38, February 2011
Available online http://www.academicjournals.org/JSTER
ISSN 2141-6559 ©2011 Academic Journals

Full Length Research Paper

Revisiting higher education data analysis: A Bayesian perspective

M. Subbiah[1]*, M. R. Srinivasan[2] and S. Shanthi[3]

[1]Department of Mathematics, K.C.G College of Technology, K.C.G. Nagar, Karapakkam, Chennai - 97, India.
[2]Department of Statistics, University of Madras, Chennai, India.
[3]Department of Computer Science and Engineering, Rajalakshmi Institute of Technology. Chennai, India.

Accepted 01 November, 2010

There has been an increasing interest by the educationists and policy makers alike on the delivery and working of educational system at all levels. Despite the extensive attention devoted to the empirical analysis of the data, most of the statistical techniques used in evaluating the data do not take account the qualitative nature of the data. The solicitation of expert's opinion and dealing with subjective probabilities are dealt with in a very limited way for analyzing educational data through Bayesian approach unlike other areas like industry and management. In this paper an attempt has been made to outline the advantages of the Bayesian approach for analyzing uncertainties involved in education data management that could be more appropriate for academic planning. The scheme has been well tested and exemplified through University student's enrollment data to underline procedure and relevance of Bayesian scheme in a multiple regression model.

Key words: Bayesian estimation, education data, expert opinion, predictive models, prior probability.

INTRODUCTION

In most recent times, higher educational institutions are not only limited to education delivery but they have to emphasize on financing and management including access, relevance, values and ethics, quality education and assessment and marketing. In addition, educational institutes are also subject to the vagaries of market forces due to stiff competition and demanding customers like students and corporate. Also, an institution must decide how to utilize the available resources to best achieve their goals and objectives in terms of creating a demand for a good service, attract students, awareness of the available products, and marketing the targeted potential customers. The challenges faced by the administrators include recruitment of students with certain characteristics such as academic ability, ethnic diversity, and qualifying status of the enrolling student community. These types of goals are increasingly difficult to achieve in recent times as the growing competitiveness among institutions, possibilities of creating diversified inter-disciplinary courses and growing demands of the student's perception about job market. In achieving the goals of present day education system, it becomes imperative to learn from the past, understand the present status, and make decisions based on reasonable prediction or forecasting of future events.

Therefore, a data driven analysis together with elicited expert's opinion is required for the best use of resources in order to realize the goals. A growing realization of the need to examine data pertains to educational and pedagogical system in greater detail has been identified. In recent times, there has been a dramatic upsurge in the collection and use of educational data which provides ample scope of modeling with the available large scale inter dependent data and the research has been directed towards similar technologies (Giesbers et al., 2007; Daphne et al., 2009; Johnson et al., 2009).

Baker and Yacef (2009) have reviewed the history and current trends in educational data mining and have discussed the increased emphasis on prediction, the emergence of work using existing models to make

*Corresponding author. E-mail: sisufive@gmail.com. Tel: 044-2450 3232. Fax: 044-2450 2898.

scientific discoveries. Romero and Ventura (2007) have identified that the educational data mining as a young research area and emphasized the necessity for more specialized and oriented work on educational domain in order to obtain successful applications.

The evaluation of teaching in education management has been considered as more relevant measure in most of the policy making decisions involving faculties such as promotion, salary etc. In many cases, examination results have been analyzed periodically by the education administrators (Goldstein et al., 1993; Ramaswami and Bhaskaran, 2010) and the outcomes could be associated in evaluating teaching effectiveness and performance.

Vialardi et al. (2009) have analyzed date pertain to course enrollment taking into account of academic performances and similar characteristics and have made recommendations to support students in their choice of appropriate courses. Predictive models for performance of students (Bekele and Menzel, 2005; Hien and Haddway, 2007) based on student's perception have been discussed in the literature. In Bayesian perspective, Zwick (1993) has analyzed the factors to predict the first-year average and final grade-point average in research programs.

However, the implications of non-inclusion of many factors involved in analyzing overall parameters could lead to a misconception while interpreting the results. It has been observed that an appropriate method that could incorporate subjective nature of the available information in educational data would be an added advantage in dealing with the uncertainties involved in these processes. A typical Bayesian analysis would be more relevant and need for such data analysis through a properly devised set of priors in the form of suitable and plausible probability distributions.

The purpose of this paper is to outline the advantages of the Bayesian approach for analyzing uncertainties involved and to incorporate the elicited information with available data that could be more appropriate in analyzing education management objectives. More specifically, one of the advantages of Bayesian analysis in terms of incorporating available evidence and opinion (Spiegelhalter, 2004) has been investigated. The subtle yet important difference between the two paradigms of statistics has been extensively used in terms of reasonable prior elicitation, efficient way of data handling, and the interpretation of statistical results to suit the multifaceted education administrative goals. This provides an ample scope to envisage a detailed statistical analysis which exploits the background information together with sample data. The study makes an attempt with a predictive model approach using regression analysis for the enrollment data that include factors such as qualification, gender obtained from the department of statistics, University of Madras, Chennai, India. Model parameters have been estimated in the light of information borrowed from the opinions of experienced department resources, the pattern of student's admissions over the period, and the present enrollment data.

BAYESIAN STATISTICS

The Bayesian methods essentially do not make distinction between model parameter (θ) and data (x). Both are considered as random variables so that 'data' are observed variables and 'parameters' are unobserved variables.

The main aim of Bayesian inference is to calculate the posterior distribution of the parameters, which is the conditional distribution of parameters given the data. The uncertainty on the parameter θ could be modeled through a probability distribution π, called prior distribution. The inference is then based on the distribution of θ conditional on x, p(θ/x) called posterior distribution obtained using Bayes' theorem.

Hence, the mechanism of the Bayesian approach to make inference consists of three basic steps; (i) assign priors to all unknown parameters, (ii) define the likelihood of the data given the parameters and (iii) determine the posterior distribution of the parameters given the data using Bayes' theorem. The first step remains a major stumbling block in the Bayesian process. The usual questions being raised are defining suitable models and constructing appropriate probability specifications.

The advent of computers and proved techniques available for computing the conditional probability distribution and in evaluating the model, answers the questions raised earlier to a greater degree. In particular, the much feared dependence of conclusions on subjective, prior distribution could also be examined and explored. The ability to include prior information in the model is not only an attractive pragmatic feature of the Bayesian approach and is theoretically vital for a guaranteed coherent inference.

The way of expressing the beliefs about θ is by taking into accounts both prior beliefs and the data. Though the prior beliefs may differ, there may be a common agreement on the way in which the data are related to θ. This eventually reflect in posterior but will turn out that if enough data is collected, then the posteriors will usually become very close and then Bayes theorem encapsulates the technical core of Bayesian inference.

However, practical problems in statistics include several parameters of interest and conclusions will often be drawn on one or more parameters at a time. Continuing advances in computing mean that many analyses previously considered computationally hopeless can be handled quite easily, even without access to large mainframe computers. Monte Carlo methods are ideally suited for the task of passing many models over one data set in Bayesian methods. Practitioners are increasingly turning to Bayesian methods for the analysis of complicated statistical models and this move seems due in large part to the advent of inexpensive high speed computers and rapid development of stochastic integration methodology, especially Markov chain Monte Carlo (MCMC) approaches such as Gibbs sampler (Gelman and Rubin, 1996; Hobbs, 1997; Gelman et al., 2003; Dongen, 2006).

In this paper, a reasonable elicitation process has been attempted to understand the student's enrollment and admission process in the statistics department, University of Madras, India and subsequently subjective priors are assigned for the parameters in the predictive model. This elicitation of opinion is an important and crucial step for making subjective Bayesian inference and computations for obtaining posterior summary have been implemented with open source software WinBUGS that could be accessed from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

MATERIALS AND METHODS

The study makes use of enrollment data available with the department of statistics, University of Madras, India which offers post graduate programs in statistics and actuarial science. The eligibility criteria for admission into statistics program includes the under graduate degrees in mathematics, statistics or three degree course with at least mathematics or statistics as one of the subjects; and for program on actuarial science, any degree with 85% marks in school completion mathematics course.

The admission process includes scrutinizing the qualification criteria eligibility for each of the applicant, entrance examination on specific topics of mathematics and statistics and ranking the applications based on the cumulative score of both entrance and qualifying examinations adhering the government norms and guidelines for admission in to any academic programs.

However, the performance in the entrance examination could highly be influenced by the qualification attributes like consolidated marks and the nature of the basic degree of the applicants. Added to these, gender of the candidates has been considered for the analysis as this would also have role to play in the seat sharing matrix. This factor inclusion needs a careful intervention with the department authorities to understand the trend in enrollment data over the years that could be more essential and reasonable in any Bayesian analysis.

The initial data analysis with the enrollment data of the academic year 2010 to 2011 indicates the demographic details of the applicants and the details may not be so essential to include in this paper. Subsequently the entrance examination performance has also been analyzed to understand more critically whether the academic output expected from the small and medium scale enterprises (SMEs) have some correlation with the marks obtained by the applicants. This way of extracting information is quite meaningful in constructing plausible priors in to the analysis that could incorporate the beliefs or opinions of domain and subject matter experts. Table 2 provides a sample of analysis yet more important indicator for identifying the perception for setting questions in such a competitive examination.

From Table 2, it could be observed that the nature of each topic and the base set by the SMEs are reflected to a better extent; that is, more evident in the result which deals with questions to test the ability of data interpretation skills with lesser mathematical details. This would enable to understand the performance trait of students who have the aspiration to join in the respective courses; however, beyond such summary statistics, the information acquired will be more essential to strengthen the general beliefs about the system. Such accrued knowledge of expert's opinion is particularly interesting and offers useful information to others and the present analysis exploits to the application of multiple linear regression model.

A multiple regression is a statistical technique that might describe and model the relationship between variables that involves more than one regressor and a response variable and the general mathematical form the model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...+ \beta_k x_k + \varepsilon$$

Where y denotes the response, x_i denotes the regressors and an error term ε which is generally assumed as a random variable with mean 0 and variance σ^2. The term linear is used to indicate a linear function of the unknown parameters β's and Gelman and Hill (2007) provides details of linear regression analysis including Bayesian estimators.

In general, the prior information about β's could be described by p-variate (p = k + 1) normal distribution with a mean vector and a covariance matrix and this could be mostly act as a vague prior. However, if the data analyst could make explicit statements about the form of the priors for β's then the statistics involved in the analysis could easily be understood and this is possible when the meaning of β's clearly defined. Such non-informative or mathematically tractable conjugate priors may not completely represent the background information pertain to the domain of investigation and the background information and /or expert's opinion could suitably be postulated as plausible prior distribution.

The explicit form of the regression model used in this analysis involve regressors related to educational indicator variables such as qualification mark (Q), gender (G) of the applicants and the type of the qualifying degree. Of these, Q is the marks in their undergraduate programs and calculated as percentage; G is a categorical variable with two levels boys and girls; qualifying degree (D) is also categorical with four levels: maths major (M), statistics major (S), three major (T), and any other (O) programs such as commerce with actuarial specialization, or engineering faculties; scores of entrance examination (E) is considered as response variable.

The underlying model would be E ~ normal (mu, σ^2) and mu = $\beta_0 + \beta_1 Q + \beta_2 G + \beta_3 M + \beta_4 S + \beta_5 T$ with reasonable priors for β's. The present study has incorporated the meaning of parameters β's in the view of overall admission pattern over a period, the general impact of qualifying degree and gender in the performance of entrance examination, in the process of estimating the parameters β's**.** Hence, before constructing the priors for the Bayesian analysis, it would be necessary to understand the data handling for two categorical variables G and D to account for the effect that the variables may have on the response. This is done by the use of indicator variables as 0 and 1 to identify the gender of the applicant as a girl or a boy respectively. Similarly the four levels of factor D can be modeled by three indicator variables M, S, and T, are shown in Table 1.

Thus, β_3 , β_4 and β_5 measures the effect of maths major, statistics major, and three major type of qualifying degrees respectively, compared to other faculties as these three major play significant roles in the course curriculum than the rest. Other effects such as the relative effect of maths major compared to statistics major can also be derived by making a direct comparison of the corresponding β's and similar interpretation could be applied to β_2, which measures the effect of the performance traits based on gender.

In view of the above meaning and the mathematical structure of the model, the absence of qualification marks for a girl student

Table 1. Four levels of factor D modeled by three variables M, S, and T.

Type of Degree	M	S	T
Other	0	0	0
Maths major	1	0	0
Statistics major	0	1	0
Three major	0	0	1

Table 2. Topic wise summary of entrance examination conducted for the admission of two postgraduate courses; statistics (Stat) and actuarial science (Acts).

Topic No.	No. of appearance		No. of attempts		No. of correct answers		Percentage	
	Stat	Acts	Stat	Acts	Stat	Acts	Stat	Acts
1	197	218	193	216	92	95	47.67	43.98
2	192	223	190	220	47	52	24.74	23.64
3	208	236	203	233	88	104	43.35	44.64
4	180	234	174	232	46	60	26.44	25.86
5	203	214	198	213	70	61	35.35	28.64
6	190	225	185	222	54	63	29.19	28.38

belong to other stream of qualifying degree (all x's = 0), a non-informative prior could be assigned to the intercept β_0. Based on a general belief that a positive correlation exist between Q and E, prior to the intercept β_0, could be considered as normal with mean and variance as 10^{-4}.

The coefficient β_2 can be understood as a mean difference of entrance scores between boys (E_M) and girls (E_G) as $E_M = \beta_0 + \beta_2$ (for boys, $x_2 = 1$ and other x's = 0) and $E_G = \beta_0$. The background information is effectively used to obtain the prior for β_2, that on an average girls score are in general higher than that of boys as the department admission in the past few years has been dominated highly by girls; more importantly the admission for the course is solely decided by the top performance of entrance examination and the qualification marks so that the rationale for this belief could be converted in to a plausible prior as normal with a mean of -0.5 and a variance of 1.

Further, β_3 the coefficient of M, could be considered from the belief that on an average the students from maths group could perform 20% better than other (O) group of students and the scores are measured in percentages the prior for β_3 has been taken as normal with a mean of 0.2 and a variance of 1. Similar notion could be followed for other two β's but the little advantage of three major (which includes maths and statistics) has also been considered so that β_4 is taken as normal with a mean of 0.2 and a variance 1 and β_5 has been taken as normal with a mean 0.4 and a variance 1. Also, prior for the variance parameter in the model is a typical inverse-gamma distribution with parameters 1 and 3 (Subbiah et al., 2008).

RESULTS

The flexibility in the inclusion of educational domain details for estimating and interpreting the parameters in the regression model has been identified as strength and the need to apply Bayesian analysis. The analysis that is based on the likelihood function and prior distribution to fit the Bayesian regression model with appropriate priors has been implemented in WinBUGS. The data input and other MCMC requirements have been carried out systematically and the script enabled option in WinBUGS makes the computations relatively easier. The posterior summary for the required parameters are presented in Table 3 and in obtaining a α-posterior credible interval for a parameter say θ, the interval (LL, UL) has been considered where LL is a (1 - α)/2 quantile and UL is a 1- (1 - α)/2 quantile for the posterior distribution of θ.

The mean values presented in the table is the corresponding estimated values for the parameters that may be considered as the relative importance in explaining the variation in the response variable viz, entrance examination marks. For example, a negative value for the gender indicates the belief that average score of girls would be higher than that of boys; similarly other parameters with respective estimated values support the prior beliefs except the marginal difference in the comparison among the students of statistics and other faculties of study.

Apart from the listed parameters included in the model, deviance (Observed – Fitted) is also calculated as a part of regression analysis and the extensive results have not been included in Table 3 due to paucity of space. However, the relevant calculation with the deviance has been carried out to have a graphical analysis of residuals to investigate the adequacy of the fit of the regression model and to check the underlying assumptions. In the normal probability plot, only a small departure from normality has been observed that in general do not affect the model greatly and as a test for outliers using inlying score plot indicates no extreme observations in the data

Table 3. Posterior summary (mean and standard deviation – SD) of the parameters in the linear regression model with lower (LL) and upper (UL) limits of posterior intervals.

Parameters	Mean	SD	LL	UL
Constant	-1.143	6.789	-14.43	12.2
Quantile. Mark	0.4965	0.09675	0.3065	0.6852
Gender	-0.2198	0.9178	-2.021	1.566
Maths	0.5285	0.9217	-1.279	2.338
Statistics	-0.4037	0.9177	-2.2	1.395
3 Major	0.5281	0.9672	-1.371	2.431
σ^2	103.7	15.82	77.08	138.9

set except for only two values numbered at 7 (E: 50, Q: 52, boy, maths) and 24 (E: 56, Q: 56, girl, statistics).

Also, as one of the essential steps in MCMC, Figure 1 displays the convergence diagnostic graphs to assess whether Markov chain has converged. The kernel density plot for the parameters shows the smooth, unimodal shape of posterior distribution for the parameters. Also, the Gelman-Rubin statistic has been provided as it is widely used recommendations for monitoring convergence of a multiple chains includes the examination of time series graphs of simulated sequences of each parameters of interest (Gelman et al., 2003). The present analysis has used three parallel chains and found that convergence criteria of Gelman-Rubin test value have been satisfied by all the parameters.

Further, the entire data set is divided into two parts based on non-repeated random generator to pick 50% of the observation so that one set is used to fit the model and the other one is considered to validate the model. This yields a very similar estimated values for the parameters as listed in Table 2 and the residual analysis also indicates no deviation in model fit. Also, to study the performance of the method, some Monte Carlo simulation has been carried out to generate 1000 samples from the appropriate underlying distributions to estimate parameters using the Bayesian method and summary of results has shown a reasonable model fit based on the diagnosis methods.

Conclusion

Most of the educational data possess the qualitative nature or perception based information where the opinions are recorded. As illustrative case, the ability and interest of a student on different subject areas could be drawn as meaningful priors for levels of interest combined with the information provided by experts in shaping course choice for the student. This could further be extended to the predictive analysis for the student as a monitoring, preventive and corrective mechanism that would be updated periodically through the performance recorded through their examinations. Also, predictive models could appropriately be constructed for a futuristic plan based on the student's attitude and characteristics, job market information and potential factors for investment.

Bayesian analysis can effectively use subjective and useful evidence to make and update the information pertaining to educational data. For example, information of expert's opinion could be built into the Bayesian model to define optimal strategies for education data analysis. Bayesian analysis could also be used to track and quantify additional information that would influence the decisions through properly elicited prior sets. Most of the theoretical work lies in choosing an appropriate prior distribution based on prior information and conditions and obtain a well behaved posterior distribution.

Situations may warrant avoiding theoretical considerations entirely and chose a 'subjective' prior distribution representing, at best, the scientific knowledge about the set of uncertain parameters. In practice, however, subjective knowledge is hard to specify precisely, and so it is important to study the sensitivity of posterior inference (Kadane and Wolfson, 1998). Bayesian methods could incorporate diverse sources of information, including subjective opinions, historical observations and model outputs that represent the educational data in a more pragmatic way.

Investigations that incorporate beliefs and expertise together with relevant data do provide meaningful interpretation could be helpful to make reasonable judgments for future academic course of action. Rubin (1983) has pointed out that the role of Bayesians is to think about parametric structure and work towards enhancing model. In this paper an attempt has been made to use the institution enrollment data together with the appropriately derived information from domain experts to perform Bayesian regression analysis. The present work emphasizes the need to device plausible priors for the parameters that directly provide a meaningful interpretation to understand the relevance of educational indicators involved in the study. Many studies have been attempting to determine various factors that affect academic success yet educational data provide plenty of opportunities for effective Bayesian analysis that collaborate experts of

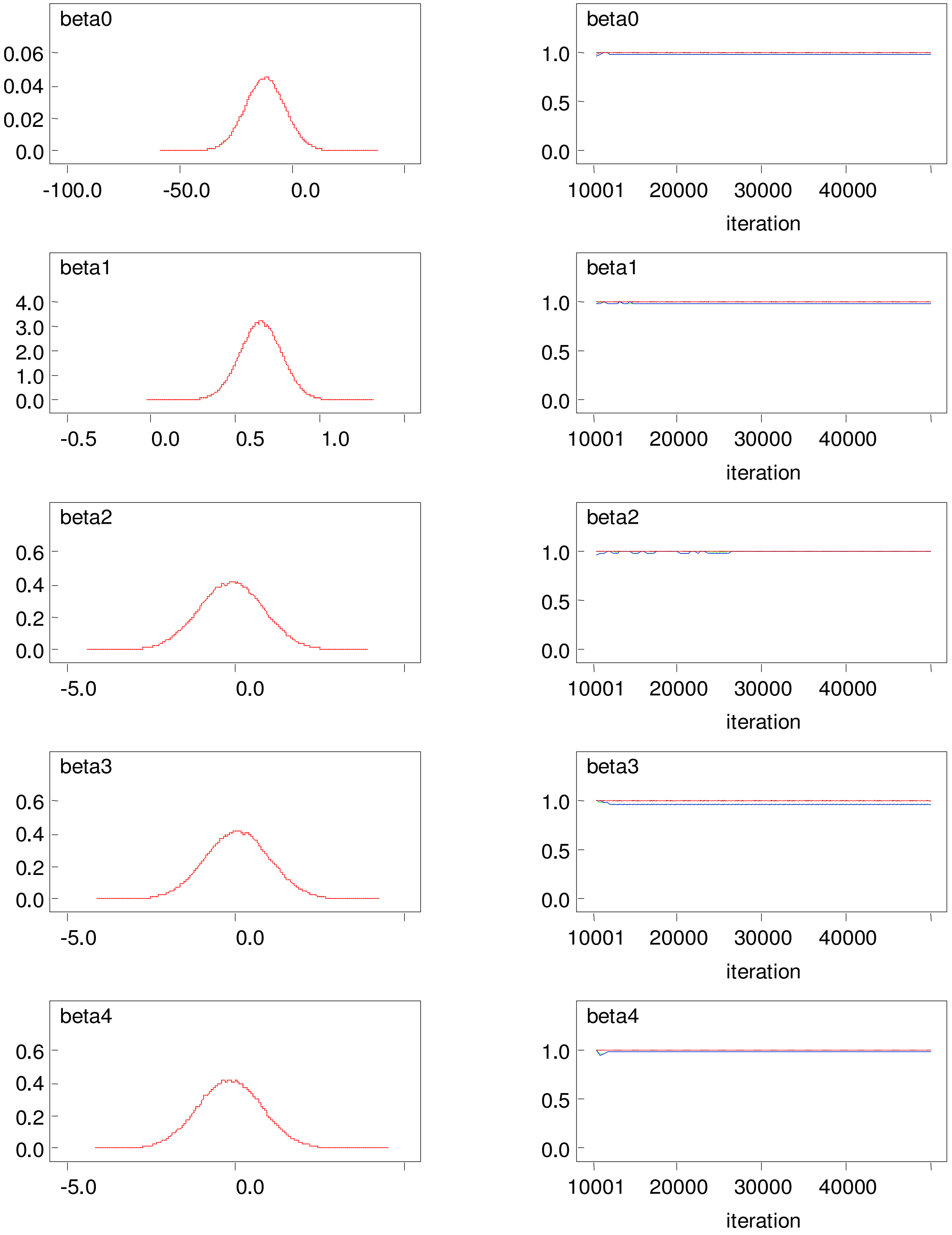

Figure 1. Diagnostic plots and Gelman – Rubin test for monitoring the convergence for each of the model parameters.

both in statistics and in education management.

Further, data mining has become a handy tool among educational research communities that mainly emphasizes analyzing data based on traditional education systems, web-based courses, learning content management system, school effectiveness, student's performance and faculty performance evaluation. Hence, there is a need for the research groups to consider the highly subjective but more relevant information available in the educational data and the corresponding analysis and building models.

While data collection through properly framed questionnaires, the elicited expert's opinion should also be placed in an appropriate method so that conclusions and interpretation of results would lead to more reasonable and pragmatic decisions in education management. The present work could be considered as a necessary extension in an effort to build a model that is more suitable for these classes of data sets and to consolidate various operational components of higher education institutions; for devising more plausible statistical models to analyze and interpret the available data to make relevant decision support systems. Bayesian analysis could be an attractive and a more suitable statistical procedure to classify, analyze and build predictive models based on derived information from suitable sources and appropriately incorporates in the process of educational management data analysis.

REFERENCE

Baker RSJD, Yacef K (2009). The State of Educational Data Mining in 2009: A Review and Future Visions, J. Educ. Data Mining. http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMVol 1 Issue 1_BakerYacef.pdf

Bekele R, Menzel W (2005). A Bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students. Proc. International Conference on Artificial Intelligence and Applications.

Daphne P, Tan GSH, Ragupathi K, Booluck K, Roop R, Ip YK (2009). Profiling Teacher/Teaching Using Descriptors Derived from Qualitative Feedback: Formative and Summative Applications. Res. Higher Educ., 50: 73-100.

Dongen SV (2006). Prior specification in Bayesian statistics: Three cautionary tales. J. Theoretical Biol., 242: 90-100.

Gelman A, Hill J (2007). Data Analysis Using Regression and Multilevel / Hierarchical Models. Cambridge University Press.

Gelman A, Rubin DB (1996). Markov chain Monte Carlo methods in biostatistics. Statistical Methods in Medical Research, 5, 339-355

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). Bayesian Data Analysis. Chapman & Hall/CRC.

Giesbers B, Bruggen JV, Hermans H, Brinke B, Burgers J, Koper R, Latour I (2007). Towards a methodology for educational modelling: a case in educational assessment. Educ. Technol. Soc., 10: 237-247.

Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, Nuttall D, Thomas S (1993). A Multilevel Analysis of School Examination Results. Oxford Rev. Educ., 19: 425-433.

Hien NTN, Haddawy P (2007). A Decision Support System for Evaluating International Student Applications. Proceedings of 37th ASEE/IEEE Frontiers in Education Conference.

Hobbs BF (1997). Bayesian Methods for analyzing Climate change and water resource uncertainties. J. Environ. Manage., 49: 53-72.

Johnson BG, Phillips F, Chase LG (2009). An intelligent tutoring system for the accounting cycle: Enhancing textbook homework with artificial intelligence. J. Acc. Educ., 27: 30–39.

Kadane JB, Wolfson LJ (1998), Experiences in Elicitation, J. Royal Stat. Society. Series D (The Statistician), 47: 3-19

Ramaswami M Bhaskaran R (2010). A CHAID Based Performance Prediction Model in Educational Data Mining. Int. J. Comp. Sci., 7: 10 - 18.

Romero C, Ventura S (2007). Educational data mining: A survey from 1995 to 2005. Expert Syst. Appl., 33: 135–146

Rubin DB (1983). Some applications of Bayesian Statistics to Educational data. Statistician, 32: 55 – 68.

Spiegelhalter DJ (2004). Incorporating Bayesian ideas into health-care evaluation. Stat. Sci., 19: 156-174.

Subbiah M, Kishore Kumar B, Srinivasan MR (2008). Bayesian Approach to Multicentre Sparse Data, Communication in Statistics - Simulation Comput., 37: 687 — 696

Vialardi CB, Shafti JL, Ortigosa A (2009). Recommendation in Higher Education using Data Mining Techniques, Educational Data Mining. http://www.educationaldatamining.org/EDM2009/uploads/proceedings/vialardi.pdf

Zwick R (1993). The validity of the GMAT for the prediction of grades in doctoral study in business and management: An Empirical Bayes approach, J. Educ. Behav. Stat., pp. 91-107.