Bayesian Model Averaging:- An Application in Cancer Clinical Trial

Atanu Bhattacharjee[*]

[*]Malabar Cancer Centre, atanustat@gmail.com

Bayesian Model Averaging:- An Application in Cancer Clinical Trial

Atanu Bhattacharjee

Abstract

Data driven conclusion is mostly accepted approach in any medical research problem. In case of limited knowledge of deep idea about supportive data on the problem, automatic digging of the variable plays important role for insight view of the study. Bayesian model averaging can be considered for automatics variable selection. It can be used as an alternative of stepwise regression method. The aim of this paper is to show the application of Bayesian modeling averaging in medical research particularly in cancer trial. Method is illustrated on Bone marrow transplant data. It can be recommended that BMA can be used frequently in data selection and as a tool of exploratory data analysis method. It is very handy method of choice for data analysis.

Bayesian Model Averaging:- An Application in Cancer Clinical Trial

Atanu Bhattacharjee

Lecturer (Biostatistics)

Division of Clinical Research and Biostatistics

Malabar Cancer Centre, Thalassery, Kerala-670103.

Abstract

Data driven conclusion is mostly accepted approach in any medical research problem. In case of limited knowledge of deep idea about supportive data on the problem, automatic digging of the variable plays important role for insight view of the study. Bayesian model averaging can be considered for automatics variable selection. It can be used as an alternative of stepwise regression method. The aim of this paper is to show the application of Bayesian modeling averaging in medical research particularly in cancer trial. Method is illustrated on Bone marrow transplant data. It can be recommended that BMA can be used frequently in data selection and as a tool of exploratory data analysis method. It is very handy method of choice for data analysis.

Keywords:- Prior, Posterior , Bayesian Information Criteria and Decision Information Criteria.

Bayesian Model Averaging:- An Application in Cancer Clinical Trial

In case of limited subject matter information, automatic variable selection is useful to explore relation between outcome and covariates. Although, logistic regression and stepwise selection are general approach to be considered to explored the relation between outcome and covariates. But it is tedious and time consuming. But the resulting estimates is depends on the choice of the selected model. The particular selected model can leads to the significance or insignificance test result. In addition, selected regression model do not show the uncertainty about choice of the assumed model in the Statistical Interpretation.

Bayesian methods in clinical trial and in medical research have become pretty well-known in last ten years due to flexible application, easy interpretation, good operative features. Bayesian is more useful in longitudinal, survival and clinical drug treatment comparison. It is performed quite well in adaptive trial design.

In model selection it can also be useful and can be choice as automatic model selection. Bayesian Modeling Averaging (BMA) givs the lack of confidence in a canonical regression. Let y be the dependents variable, α_γ a constant, β_γ, the coefficients, and ε a normal IID having variance σ^2.

$$Y = \alpha_\gamma + \beta_\gamma X_\gamma + \varepsilon, \varepsilon \sim N(0, \sigma^2 I). \qquad (1)$$

Based on primary end point of the study it is easy to define Y. But among several$X \epsilon x$, it is not possible to select best X. The problem in variable selection is X not Y.

BMA attempts the problem through estimating the models among all probable combination of $\{X\}$ and gives the information about weighted average of all of them. Let X is represented as $X_1, X_2, X_3$and X_4. Then possible models are $2^4 = 16$.

In experimental study, K numbers of variables may be considered. The number of possible model would be 2^k. But, selection of best model among 2^k models is very difficult. BA with weighted average solve the problem.

Modeling

Let the Model M_γ , prefers the covariates X_γ. Then by Bayes theorem it can be presented as

$$p(M_\gamma/Y, X) = \frac{p(Y/M_\gamma)p(M_\gamma)}{p(Y/X)} = \frac{P(Y/M_\gamma, X)p(M_\gamma)}{\sum_{s=1}^{2^k} P(Y/M_\gamma, X)p(M_\gamma)} \qquad (2)$$

Now, the likelihood of $p(Y/X)$ for the particular data set will not change. Then it can be stated that $P(Y/M_\gamma, X)$ is the proportional to the marginal likelihood of $P(Y/M_\gamma, X)$ time prior probability $p(M_\gamma)$. The term $P(Y/M_\gamma, X)$ is the posterior model probability of the data given model.

Let the coefficient(β) in equation (1) is estimated by $\hat{\beta}$. Then

$$p(\hat{\beta}/y, X) = \sum_{s=1}^{2^k} p(\hat{\beta}/M_\gamma, Y, X)p(M_\gamma/X, Y). \qquad (3)$$

Here, s is the possible set of all models. The prior probability models i.e. $p(M_\gamma) \propto 1$ to gives the minimum effect of knowledge.

Acceptance of BMA

In classical approach, $\alpha = 0.05$ is well accepted principle in the scientific community. BMA can be used in this manner. A 95% threshold value can be applied for posterior estimation. In this stydy, 95% posterior threshold is used for selecting best predictor. The analysis is performed to compare the performance of linear regression selection step and BMA. The well-know and widely, applied method in stepwise regression is found poorly performed in the theory and case –studies [Raftery et al., 1997 and Malek et al.,2007].

Data Methodology

The graft dataset comes from a non-depleted allogeneic bone marrow transplant from an HLA-identical sibling donor for a haematological malignancy of 37 patients in GvHD trial. Information was recorded on recipient age, donor age, type of leukemia and status of pregnancy. The aim of the work by Bagot et al,. [1988] was to detect the donor/recipient pairs at high risk of GvHD who might benefit of bone marrow T-cell depletion. The basic characteristics of the data are provided into the Table 1 and Table respectively.

Bayesian Model

Bayesian methodology in statistical Inference is well-documented [Gelman et a., 2004, Carlin and Louis 2008 and Ibrahim et al., 2001]. Bayesian differs from the classical in terms f uncertainty of the unknown parameter in a model is expressed through an entire distribution called the prior distribution. The term "prior" expressed the uncertainty of the parameters before collecting the data in the study. The main inferential tool is possible through posterior distribution by application of Bayes Theorem. If $\pi(\hat{\beta})$ be the prior information about parameter β. $L(Data/\hat{\beta})$ is the likelihood function of the data given the parameter β. Then $\pi(\beta/Data)\infty L(Data/\beta)\pi(\beta)$ is posterior distribution of parameter. The major consideration in the Bayesian is the selection of prior. Priors that have a minimal influence on the overall Bayesian analysis are called non-informative prior. Flat, Vague, reference-priors are synonyms of the non-informatrive prior. The result obtained through non-informative prior is very much similar with classical approach. The non-informative prior is relatively “flat” towards distribution of the data.

Bayesian con be computed in WinBugs, SAS and R. These software packages are very powerful for data analysis. Markov Chain Monte Carlo(MCMC) is only required to performed to obtain the Bayesian results. Softwares are now a day’s compatible for Bayesian computation. MCMC is basically simulation based approach to draw sample from the posterior distribution of β and have already established as powerful tool for even complex model. Dealing with Complex model is really infeasible for classical method.

Illustration of BMA

The ‘graft data’ data set describe above having 8 variables named with “gvhd“,“index”, “dead”,“preg”, “rcpage “,“time“,“type“ and “donage “respectively. The “rcpage” is considered as outcome of interest and rest all are covariates. Now, as stated earlier a total of 28 types of model can be formed. Here, BMS library in R software is used. R is open source software and can be downloaded from www.cran-project.org. In the bms function, prior is defined as “uniform”. The results obtained through bms function on response variables are listed in Table 3.

In the above results, post mean provides the information about averaged over all model in particular coefficient. It shows that variable “gvhd” has the highest coefficient value in comparison to others. The posterior inclusion criterion gives the importance of the variable in the data. It can be noted that the PIP value and post mean are positively correlated. Only “gvhd” can be considered as an important variable in comparison to others. The posterior standard deviation gives more information: “gvhd” is positively influencing to the on response , but “rcpage” worked as negatively. The column named with “Index” gives the indices value of the variable in the data set.

Modeling

The posterior model size can be obtained by sum of all PIP. In this example, it comes around 3.60.As stated earlier, MCMC is useful for simulation techniques. The correlation between analytical observation of PMP and iteration is given in figure1. It shows that correlation between iteration and analytical observation are in decent level.A total of 2000 models are genereated. The best filled 50 models are given in figure 2.The relative performance for consideration of uniform, Fixed, PIP and random prior are given in Figure2.

Discussion

There are several procedures to detect the best fitted models. The procedures are Bayesian Information Criteria (BIC)[Neath and Cavanaugh (2012)], Decision Information Criteria [Spiegelhalter et al., (2002)] and Akkai Information

Criteria[Burnham et al. (2002)]. However, the application of those approaches only becomes useful when parameters are selected. In case of presence of crude or unselected parameters, none of the approach can be performed well. The application of BMA is very simple. The open source software R can be used to perform the BMA[Ref]. More detailed application of BMA through R are recently discussed [Clyde (2010), Raftery et al. (2010a) and Feldkircher and Zeugner (2011)].

Conclusion

The BMA is useful to explore the prior information about the prior distribution of the parameter of interest. It is useful to specify the parameters to be considered in model development stage. The stabilize prediction can only be obtained through application of BMA. An application of BMA on “graft data” set is presented. The Bayesian analysis on graft data is applied for the model illustration. It is suggested to specify about BMA in the protocol itself to develop the track for further model development in study design.

Reference

Bagot M, Mary JY, Heslan M et al. The mixed epidermal cell lymphocyte-reaction is the most predictive factor of acute graft-versus-host disease in bone marrow graft recipients. Br J Haematol 1988; 70: 403-409.

Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.), Springer-Verlag, ISBN 0-387-95364-7.

Carlin BP, Louis TA: Bayesian methods for data analysis. 3rd edition. Boca Raton, FL: Chapman & Hall/CRC; 2008.

Gelman A, Carlin JB, Stern HS, Rubin DB: Bayesian data analysis. 2nd edition.Boca Raton, FL: Chapman & Hall/CRC; 2004.

Ibrahim JG, Chen MH, Sinha D: Bayesian survival analysis. New York: Springer; 2001.

Mundry R, Nunn CL: Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat 2009, 173:119-123.

Malek MH, Berger DE, Coburn JW: On the inappropriateness of stepwise regression analysis for model building and testing. Eur J Appl Physiol 2007, 101(2):263-4, author reply 265-6.

Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. WIREs Computational Statistics 4, 199-203.

Spiegelhalter, David J.; Best, Nicola G.; Carlin, Bradley P.; van der Linde, Angelika (October 2002). "Bayesian measures of model complexity and fit (with discussion)". Journal of the Royal Statistical Society, Series B 64 (4): 583–639.

Raftery AE, Madigan D, Hoeting JA: Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association 1997, 92(437):179-191 [http://www.jstor.org/stable/2291462].

Table1: graft dataset summary of continuous variable

Parameters	Mean	sd	Parameters	Mean	Sd
Rcpage	25.43	7.50	Donage	25.81	7.83
Time	669.8	483.71	Index	2.55	2.22

Table2: graft dataset summary of categorical variable

Parameters	Status	Freq(%)	Parameters	Status	Freq(%)
Donor Pregnancy Status	Yes	10(27.02%)	Type	Acute myoloid leukarmia(AML)	11(29.72%)
	No	27(72.98%)		Acute lymphocytic leukarmia(ALL)	16(43.24%)
Dead	Yes	19(51.35%)			
	No	18 (48.64)		Chronic myeloid leukarmia(CML)	10(27.02%)

Table 3:- The coefficient table covariates on response variable.

Parameters	PIP	Post Mean	Post SD	Cond.Pos.Sign	Index
gvhd	1.0000000	12.9443355327	1.5636948959	1.0000000	6
index	0.9999999	2.2755190644	0.3000631051	1.0000000	5
dead	0.4607726	-1.0737612621	1.4614983330	0.0000000	8
preg	0.4302314	1.0833786240	1.5705677172	1.0000000	4
rcpage	0.2063585	0.0173867110	0.0541033943	1.0000000	1
time	0.1992714	0.0002177483	0.0009393523	0.8183941	7
type	0.1549690	0.0106012518	0.3709706438	0.4850185	3
donage	0.1504394	-0.0039444487	0.0358654882	0.1477814	2

Figure1:- The correlation between PMP and iterations.

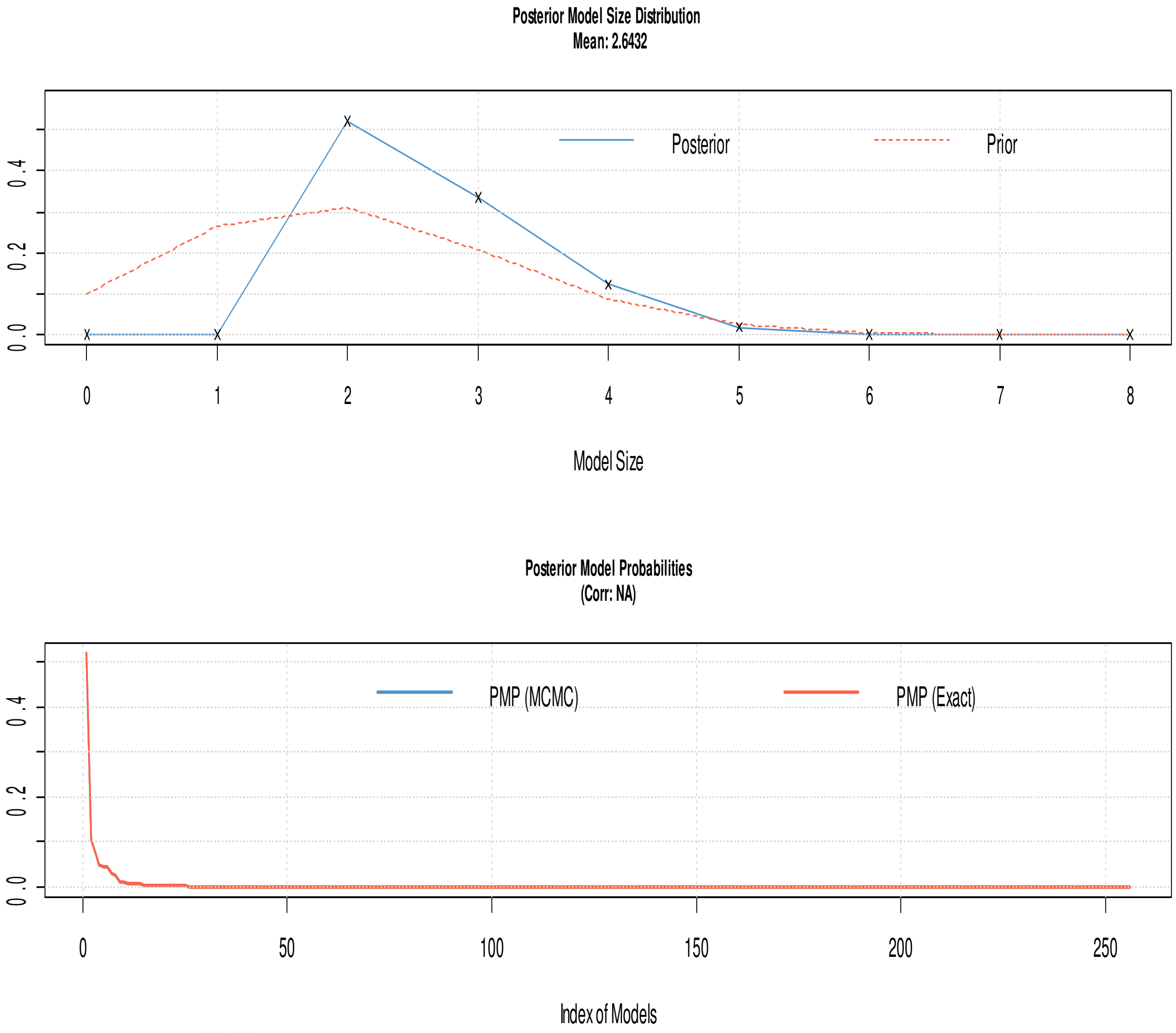
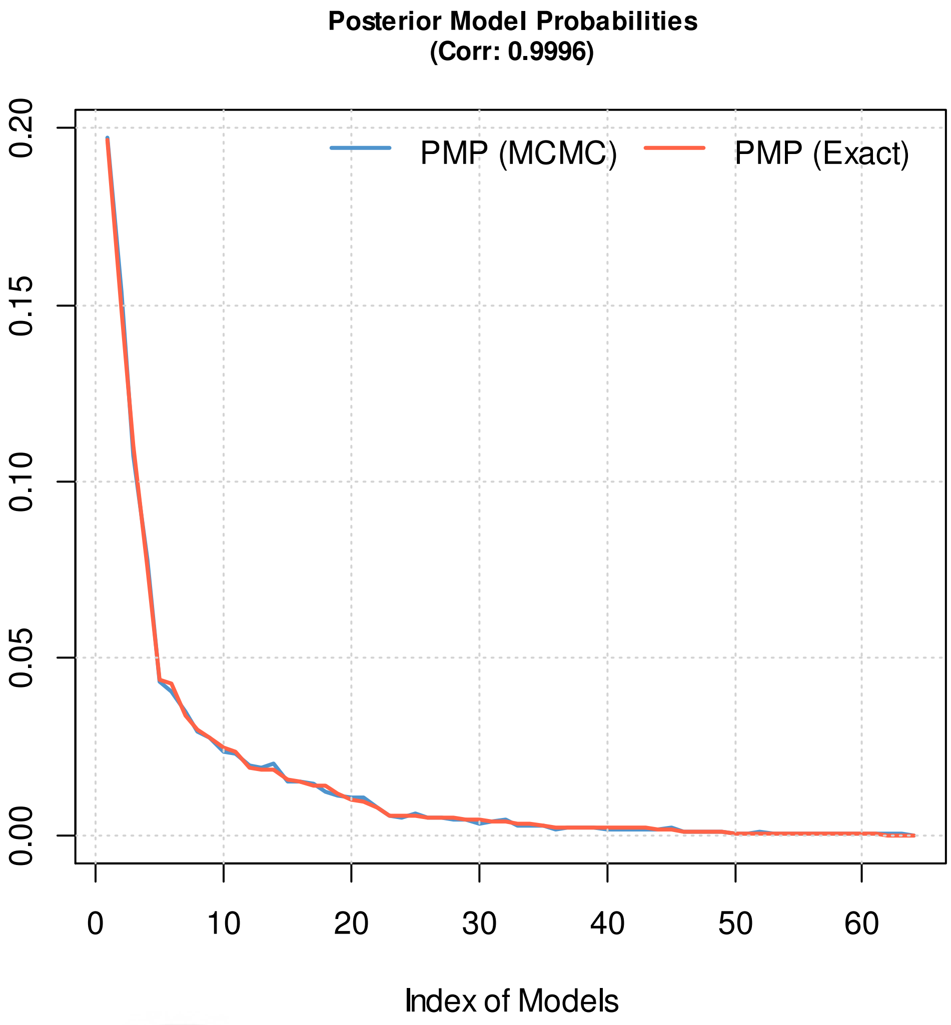


Figure2:- The best fitted models


Posterior Model Probabilities
(Corr: 0.9996)
PMP (MCMC)
PMP (Exact)
Index of Models