20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Logistic Regression and Bayesian Approaches in Modeling Acceptance of Male Circumcision in Pune, India

C. Yoo[1], A. Saxena[1], K. Krupp[1,2], V. Kulkarni[3], S. Kulkarni[3], J. D. Klausner[1,4], J. Devieux[1] , and P. Madhivanan[1,2]

[1] Robert Stempel College of Public Health & Social Work, Florida International University, Miami, USA [2] Public Health Research Institute of India, Mysore, India [3] Prayas Health Group, Pune, India [4] School of Medicine, University of California, Los Angeles, USA
Email: cyoo@fiu.edu

Abstract: Discernment analyses in survey data are being developed to help researchers better understand intentions of surveyed subjects. These models can aid in successful decision-making by allowing calculation of the likelihood of a particular outcome based on subject's known characteristics. There are many modern discernment analyses which have been used to develop predictive models in many different scientific disciplines areas. Predictive models are used in a variety of public health and medical domains. These models are constructed from observed cases, which are typically collected from various studies. The data can be pre-processed and serve as data to build statistical and machine learning models.

The most frequently used discernment analysis in epidemiological datasets with binary outcomes is logistic regression. However, modern discernment Bayesian methods — i.e., Naïve Bayes Classifier and Bayesian networks — have shown promising results, especially with datasets that have a large number of independent variables (>30). A study was conducted to review and compare these models, elucidate the advantages and disadvantages of each, and provide criteria for model selection. The two models were used for estimation of acceptance of medical male circumcision among a sample of 457 males in Pune, India on the basis of their answers to a survey that included questions on sociodemographics, HIV prevention knowledge, high-risk behaviors, and other characteristics.

Although the models demonstrated similar performance, the Bayesian methods performed better especially in predicting negative cases, i.e., subjects who did not want to undergo medical male circumcision in cross validation evaluations. Since there were less negative cases in the dataset, this indicates with smaller sample size, Bayesian methods perform better than logistic regression. Identifying models' unique characteristics — strengths as well as limitations — may help improve decision-making.

Keywords: *Logistic Regression, Bayesian networks, Discernment analyses*

1. INTRODUCTION

There are many modern discernment analyses which have been used to develop predictive models in many different scientific disciplines areas (Hastie, et al., 2001). Predictive models are used in a variety of public health and medical domains. These models are constructed from observed cases, which are typically collected from various studies. The data can be pre-processed and serve as data to build statistical and machine learning models. The most popular models in epidemiological and medical studies are logistic regression (LR), Naïve Bayes Classifier (NBC), and Bayesian networks (BN).

In this article, we show that LR, NBC, and BN share common roots in statistical learning models, and how the two models differ in predictive performance after the prediction models are built. We compare predictive performance of these three methods with data from acceptance survey of medical male circumcision among a sample of 457 males in Pune, India (Madhivanan, et al., 2011).

We show the modern discernment Bayesian methods — i.e., Naïve Bayes Classifier and Bayesian networks — have shown promising results, especially with datasets that have a large number of independent variables (> 30) and small sample size (< 30).

Logistic Regression

Logistic regression examines the relationship between a categorical outcome (dependent) variable and predictor (explanatory or independent) variables. For example, the acceptance or rejection of medical male circumcision within a specified time period might be predicted from predictor variables such as knowledge of the subject's age, knowledge of AIDS, and marital status. The outcome variables can be continuous or categorical. If $X_1, X_2,\ldots, X_n$ denote n predictor variables, and p denotes the probability of accepting medical male circumcision, the following equation describes the relationship between the predictor variables and p:

$$Log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad (1)$$

where β_0 is a constant and $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients of the predictor variables $X_1, X_2, \ldots, X_n$. The regression coefficients are estimated from the available data. The probability of accepting medical male circumcision (p) can be estimated with this equation.

Each regression coefficient describes the magnitude of the contribution of the corresponding predictor variable to the outcome. The strength of association between the predictor variables and the outcome variable is commonly measured by using the odds ratio, which represents the factor by which the odds of an outcome change for a one-unit change in the predictor variable. The odds ratio is estimated by taking the exponential of the coefficient (e.g., e^{β_1}). For example, if β_1 is the coefficient of variable X_1 ("subject is married"), and p represents the probability of accepting medical male circumcision, e^{β_1} is the odds ratio corresponding to the marriage of the subject of accepting medical male circumcision. The odds ratio in this case represents the factor by which the odds of accepting medical male circumcision increase if the subject is married and all other predictor variables remain unchanged.

Logistic regression models generally include only variables that are considered "important" in predicting an outcome. With use of P values, the importance of variables is defined in terms of the statistical significance of the coefficients for the variables. The significance criterion $P < 0.05$ is commonly used when testing for the statistical significance of variables; however, such criteria can vary depending on the amount of available data (Hosmer and Lemeshow, 2000).

Significant variables can be selected using various methods. In forward selection, variables are sequentially added to an "empty" model (ie, a model with no predictor variables) if they are found to be statistically

significant in predicting an outcome. In contrast, backward selection starts with all of the variables in the model, and the variables are removed one by one as they are found to be insignificant in predicting the outcome. The stepwise logistic regression method is a combination of these two approaches and is used to determine which variables to add to or drop from the model in a sequential fashion on the basis of statistical criteria. Although different techniques can yield different regression models, they generally work similarly. Sometimes, clinically important variables may be found to be statistically insignificant with the selection methods because their influence may be attenuated by the presence of other strong predictors. In such cases, these clinically important variables can still be included in the model irrespective of their level of statistical significance.

Bayesian Models

A Bayesian Network (BN) is a probabilistic model that contains a set of variables, known as nodes, and a set of directed links, called arcs. Directed links are nothing but conditional dependencies between variables. Altogether, nodes and arcs are called a directed acyclic graph (DAG) (Pearl, 1988)

There are many popular software such as 'Structural Modeling, Inference, and Learning Engine'(SMILE) and its 'Graphical Network Interface' (GeNIe) (Druzdzel, 2005); 'Bayesian Network Inference with Java Objects' (Banjo) (Hartemink, 2010) and bnlearn for R, which use BN learning from given data.Today, when acquiring new data is expensive and time consuming, using prior knowledge with BN learning for making predictions is becoming more common in research. A BN, also known as a belief network, is a member of the probabilistic graphical model family, which computes and represents joint probability distributions effectively over a set of variables. Here, a variable of interest can have many states. It can be continuous, categorical, a disease outcome or a hypothesis of interest. Conditional dependence or joint probability can be understood with a simple example: If an arc is drawn from node X to node Y, then in simple words X is influencing Y. Here, X is the "parent" of Y and Y is the "child" of X. An extension of this DAG may include "decedents" of Y and "ancestors" of X (Charniak, 1991).

We analyzed and learnt BN using Banjo and joint probability distribution (JPD) using GeNIe. To use data in Banjo, this dataset was first saved as tab delimited and all the variables were again converted to discrete form, with the lowest value starting from 0. This dataset was run three times each for cycles of one, two, three, and four hours in Banjo (a total of 12 times). Banjo refined or improved BN by using prior knowledge from the given data. Log likelihood values were noted at the end of each learning cycle and from the list of 12 log likelihood values, a BN having the lowest log likelihood value was chosen. Graphviz software was used to make this BN graph from the output generated by Banjo.

Keeping the graph produced by Banjo as reference, we manually recreated the BN in the GeNIe program. The dataset we used with Banjo was saved as a comma separated file and imported into GeNIe. Parameters and the likelihood value of this BN were calculated in GeNIe using its "learn parameter function." We deselected "randomize initial parameters" and kept the "confidence" as 1, and the file was saved in *.xdsl form. We then used this file and SMILE++ to calculate the predicted probabilities for each case. SMILE++ also updated the posterior distribution for the variable "accept" with all the possible permutations. This SMILE++ file/code was slightly modified to make a new column in the given dataset and assign the predicted probability for each case at the end of the row. For this calculation, we used only Markov blanket for our variable of interest. The Markov blanket of a variable A is the set consisting of the parents of A, the children of A, and the variables sharing a child with A (Jensen, 1996). The physical joint probability distribution of an event X, here "accept," can be encoded in the BN structure S as

$$p(x|\theta_s, S^h) = \prod_{i=1}^{n} p(x_i|\boldsymbol{pa}_i, \boldsymbol{\theta_i}, S^h)$$

Where $\boldsymbol{\theta_i}$ is the vector of parameters for the distribution $p(x_i|\boldsymbol{pa}_i, \boldsymbol{\theta_i}, S^h)$, θ_s is the vector of parameters$(\theta_1, \ldots, \theta_n)$, and S^h denotes the event S, for which we are factoring the physical joint probability distribution. Furthermore a node (variable) becomes independent of its non-decedents given its parents by its causal Markov condition in BN (Pearl and Verma, 1987)

2. METHODS

Data gathered from a medical male circumcision acceptability survey (n = 457) having 68 variables in total (excluding the outcome "accept") were used in this analysis. These data were discretized to make them suitable for the analysis. The outcome variable of interest was "accept'," which was dichotomous. BN and JPD were learned using Banjo and SMILE. After learning the predicted probabilities by using variables from Markov blankets and excluding marginal probabilities, sensitivity and specificity of the model were calculated by constructing an ROC curve in R (pROC package). To calculate and compare predicted probabilities of the same outcome by logistic regression (enter, forward selection, backward elimination and stepwise methods), SAS® software was used. GeNIe was used for its built in function to conduct Naïve Bayes analysis. The ROC curves, positive predictive values (PPV), and negative predictive values (NPV) of the three analysis methods are summarized and compared in Table 1.

A threefold cross-validation was performed to report the goodness of fit of the models. For logistic regression, we used SAS® software using forward selection and backward elimination method. For Bayesian network, we ran Banjo for 40 hours and selected the best Bayesian networks according to its log likelihood scoresand the best network was reconstructed in GeNIe to learn the parameters (probabilities). Also GeNIe was used to build Naïve Bayes Classifier (NBC).

For each model, we report Area under ROC curve (AUROC), positive predictive value (PPV) and negative predicted value (NPV) .

3. RESULTS

Table 1 shows the result of best LR model that was fit by using forward selection and backward elimination.

Table.1: Result of Logistic Regression model using forward selection and backward elimination.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-14.9960	181.7	0.0068	0.9342
agecat	**2**	1	-0.0621	0.4344	0.0205	0.8863
agecat	**3**	1	0.3434	0.4659	0.5434	0.4610
agecat	**4**	1	-0.0738	0.4852	0.0231	0.8792
agecat	**5**	1	-1.0066	0.5330	3.5657	0.0590
c_peruraids	**1**	1	13.1555	181.7	0.0052	0.9423
c_peruraids	**2**	1	11.4713	181.7	0.0040	0.9497
c_peruraids	**3**	1	11.8682	181.7	0.0043	0.9479
c_peruraids	**4**	1	11.5772	181.7	0.0041	0.9492
c_peruraids	**5**	1	12.1555	181.7	0.0045	0.9467
d_mythb	**1**	1	1.1492	0.3395	11.4594	0.0007
d_mythb	**2**	1	0.2364	0.2878	0.6747	0.4114
e_01_sxpt3_D	**1**	1	0.7639	0.3295	5.3728	0.0205
e_01_sxpt3_D	**2**	1	1.3629	0.5121	7.0842	0.0078
e_01_sxpt3_D	**3**	1	1.3198	0.8628	2.3398	0.1261
g_barriera	**1**	1	-1.3333	0.2506	28.3096	<.0001
g_barrierc	**1**	1	-0.6161	0.2668	5.3310	0.0209
g_barriere	**1**	1	0.7624	0.3586	4.5206	0.0335

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
g_facila	1	1	2.4681	0.5486	20.2421	<.0001
g_facild	1	1	1.3238	0.3739	12.5328	0.0004

The best Bayesian networks that is reported is shown in Figure 1. A Markov blanket of an outcome variable (accept; accepting Medical Male Circumcision) is particularly interested in Bayesian networks because once we know all the values of the variables in the Markov blanket, knowing the values of other variables that are not in the Markov blanket do not contribute in predicting the outcome. The Markov blanket of outcome variable accept is shown in Figure 2.

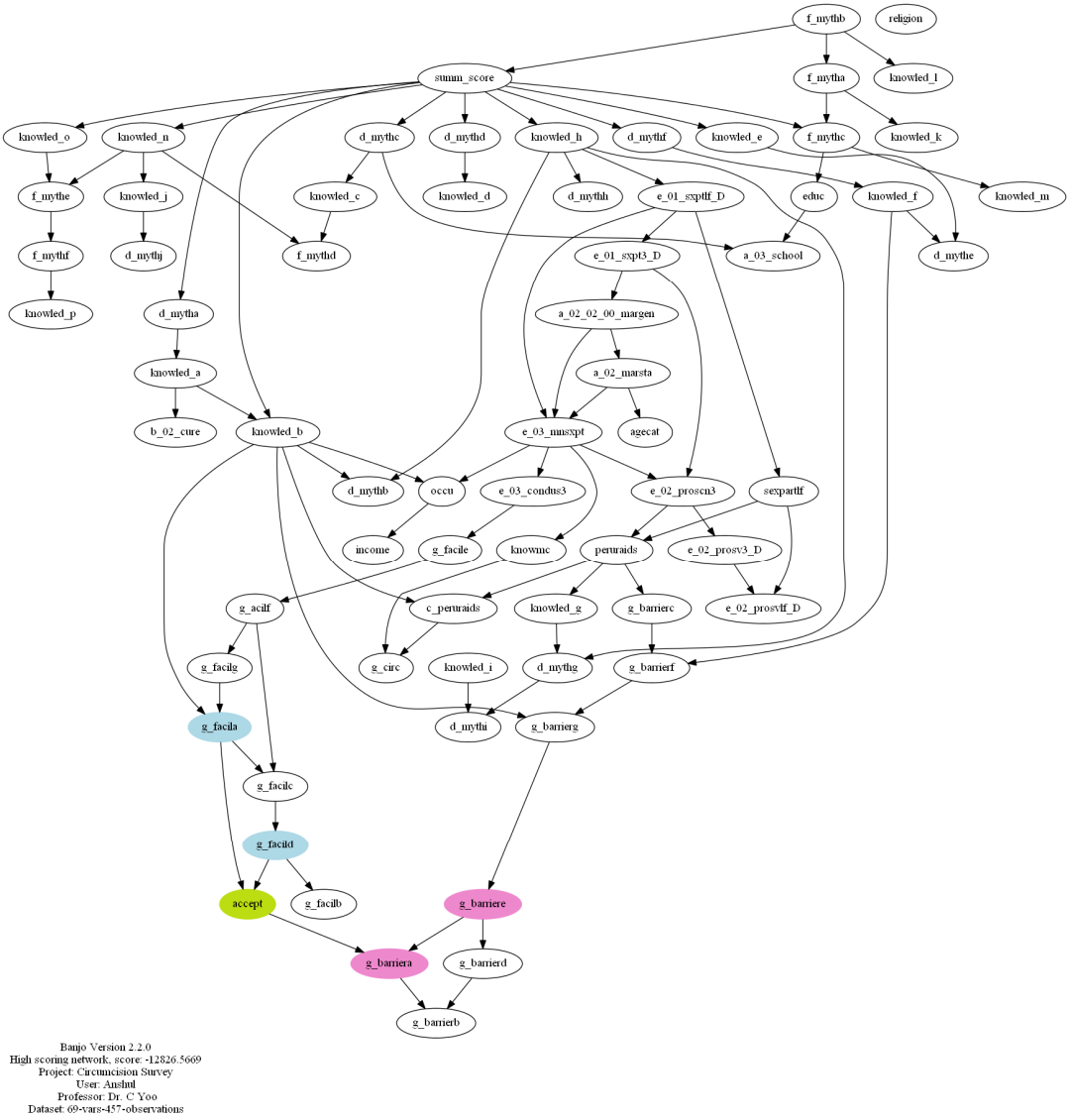

Figure 1. Highest likelihood Bayesian network given the dataset

Table 2 shows the result of three fold cross validation of Logistic Regression (LR), Naïve Bayes Classifier (NBC), and Bayesian Network (BN). Combined model (LR + BN) represents LR using BN Markov Blanket variables as predictor variables. Although they are not statistically significant, LR outperformed Bayesian methods in PPV and AUROC. Also BN outperformed other methods in NPV.

Table.2: Average Area under ROC curve (AUROC), positive predictive value (PPV) and negative predicted value (NPV) from three-fold cross validation for Logistic Regression (LR), Naïve Bayes Classifier (NBC), and Bayesian Network (BN) (standard deviation is reported in the paranthesis). Note that the AUROC, PPV, and NPV are presented as percentages.

	LR	NBC	BN	Combined (LR + BN)
PPV	71.13 (8.46)	68.33 (6.78)	66.33 (6.35)	66.83 (10.66)
NPV	62.73 (1.63)	59.67 (1.53)	80.50 (17.11)	70.50 (16.44)
AUROC	74.83 (2.99)	72.10 (2.43)	64.20 (3.50)	70.23 (1.89)

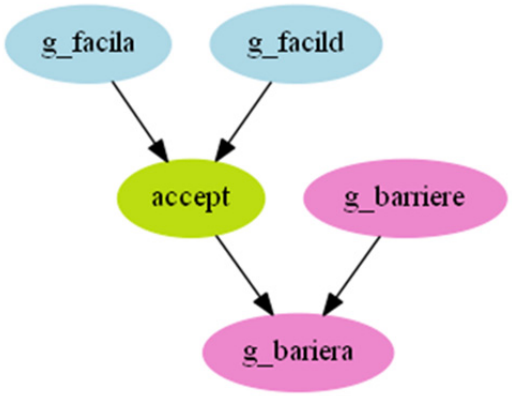

Figure 2. Interactions among the predictor variables based on the Bayesian network.

4. DISCUSSION

We have shown that Bayesian methods can be useful in helping researchers better understand intentions of surveyed subjects. Bayesian methods showed similar predicted performance compared to a widely used regression model, i.e., Logistic Regression. Moreover Bayesian methods provide ways to look into interactions among variables that will help researchers to learn causal relationships among the variables.

It will be interesting to look further into how different models' predicted performance differs by different sample size of this collected dataset. This will confirm our argument of smaller sample size, Bayesian methods will be more beneficial to use than logistic regression.

REFERENCES

Charniak, E. (1991) Bayesian networks without tears, **12**, 50–63.

Druzdzel, M.J. (2005) Intelligent decision support systems based on SMILE, *Software 2.0*, **2**, 12-33.

Hartemink, A.J. (2010) Banjo: structure learning of static and dynamic Bayesian networks.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York

Hosmer, D. and Lemeshow, S. (2000) *Applied logistic regression*. Wiley, New York.

Jensen, F.V. (1996) *An Introduction to Bayesian Networks*. Springer Verlag, New York.

Madhivanan, P., Krupp, K., Kulkarni, K., Kulkarni, S. and Klausner, J. (2011) Acceptability of Male Circumcision for HIV Prevention among high-risk men in Pune, India. , *STD*, **38**, 571.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. and Verma, T.S. (1987) The logic of representing dependencies by directed graphs. *Proceedings of AAAI*. Morgan Kaufmann, Seattle, WA, 374–379.