Journal of Mathematics and Statistics

Original Research Paper

Bayesian Network Inference in Binary Logistic Regression: A Case Study of *Salmonella sp* Bacterial Contamination on Vannamei Shrimp

Pratnya Paramitha Oktaviana and Kartika Fithriasari

Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article history
Received: 28-11-2017
Revised: 14-12-2017
Accepted: 20-12-2017

Corresponding Author:
Pratnya Paramitha Oktaviana
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Email: paramitha.oktaviana@gmail.com

Abstract: Recently binary logistic regression has been used to identify four factors or predictor variables that supposedly influence the response variable, which is testing result of *Salmonella sp* bacterial contamination on vannamei shrimp. Binary logistic regression analysis results that there are two predictor variables which is significantly affect the testing result of *Salmonella sp* bacterial contamination on vannamei shrimp, those are the testing result of *Salmonella sp* bacterial contamination on farmers hand swab and the subdistrict of vannamei shrimp ponds. Those significant predictor variables selected have been modelled in binary logit model. This paper proposes to study the statistical associations between the two significant predictor variables and the contamination of *Salmonella sp* bacterial on vannamei shrimp and to build a numerical simulation of two significant predictor variables parameters using bayesian network inference. Directed Acyclic Graph (DAG) is applied for modelling binary logit model of significant factors in bayesian network inference.

Keywords: Binary Logistic Regression, Bayesian Network, *Salmonella sp* Bacterial Contamination, Vannamei Shrimp, Parameters

Introduction

According to Hosmer and Lemeshow (2000), if there are p predictor variables, indicated by the vector $x = (x_1,x_2,...,x_p)$ and each of these variables is assumed at least interval scale, so the conditional probability could be indicated by $P(Y = 1| x) = \pi(x)$. The logistic regression model is:

$$\pi(x) = \frac{\exp(g(x))}{1+\exp g((x))} \quad (1)$$

Then the logit of that model could be written as:

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad (2)$$

If those p predictor variables are discrete or have nominal scale, the method of choice is to use dummy variables. If a nominal scaled variable has m possible values, then $m-1$ dummy variables will be needed. Suppose that the j^{th} predictor variable x_j has m_j levels. The $m_j - 1$ dummy variables will be indicated as D_{jk} and the coefficients for these dummy variables will be indicated as B_{jk}, $k = 1,2,...,m_j-1$. Then the logit of this case could be written as:

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \sum_{m=1}^{m_j-1} \beta_{jk} D_{jk} + \beta_p x_p \quad (3)$$

Binary logistic regression is a logistic regression where the response variable used is dichotomous (or it is qualitative data which has binary or two categories) and the predictor variables are polichotomous (it could be qualitative or quantitative data).

Recently binary logistic regression has been used by the researchers to identify four factors or predictor variables (X_1, X_2, X_3, X_4) that supposedly influence the response variable (Y), which is the testing result of *Salmonella sp* bacterial contamination on vannamei shrimp. This response variable (Y) has two categories: 0= if testing result of *Salmonella sp* bacterial contamination on vannamei shrimp indicate that there is no *Salmonella sp* on vannamei shrimp; 1= if testing result of *Salmonella sp* bacterial contamination on vannamei shrimp indicate that there is *Salmonella sp* on vannamei shrimp. While there are four predictor variables used: X_1: The testing result of *Salmonella sp* bacterial contamination on farmers hand swab (nominal scaled variable), X_2: The subdistrict of vannamei shrimp ponds (nominal scaled variable), X_3: The fish processing unit that supplaid by (nominal scaled variable) and X_4: The pond area in hectare (ratio scaled variable).

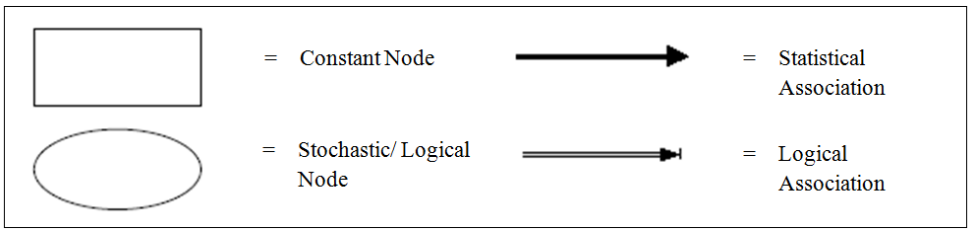

Fig. 1: The icon specification in DAG (Source: Liu, 2012)

Table 1: Link function in DAG

Link function	Formula
Identity	x_i
Logit	$\log it(p_i) = \log\left[\frac{p_i}{1-p_i}\right]$
Probit	Probit (x_i)
Cloglog	$\log(-\log(1-p_i))$
Log	$\log(x_i)$

This method obtain that there are two significant predictor variables, i.e., X_1 and X_2. Those significant predictor variables have been modelled in binary logit model.

This paper proposes to study the statistical associations between the two significant predictor variables and the contamination of *Salmonella sp* bacterial on vannamei shrimp and to build a numerical simulation of two significant predictor variables parameters using bayesian network inference. DAG is applied for modelling binary logit model of significant factors in bayesian network inference.

Bayesian Network Inference

Neapolitan (1989) in Stephenson (2000) explains that bayesian network is a specific type of graphical model, namely DAG. All of the edges in the graph are directed (the edges point in a particular direction) and there are no cycles (there is no way to start from any node and travel along a set of directed edges in the correct direction and arrive back at the starting node. The edges in bayesian network explain the joint distribution of all variables. The joint probability indicated by one set of edges can equally be indicated by another set.

Chen *et al*. (2015) explains that bayesian network is a set of variables, X and Y, that present joint probability distribution, for $i = 1,2,...,n$:

$$P(X,Y) = \prod_{x_i \in X,Y} p(x_i \mid pr(x_i)) \quad (4)$$

DAG is used to illustrate all of the parameters and variables in thae model and connect them using the edges (Liu, 2012). The icon specification in DAG is presented in Fig. 1. There are three nodes in DAG as shown as in the Fig. 1:

1. Constant Node: It is used as the icon of random variable, for example: $x_i \sim N(\mu,\sigma^2)$
2. Stochastic Node: It is used as the icon of variable which is described by the other variables, generally to predict, for example: $\mu_i = \beta_0 + \beta_1 x_i$
3. Logical Node: It is used as the icon of observation value, hyper-parameters or constant, for example: $N = 100$, x_i

Link function in DAG is presented in Table 1.

Analysis and Result

Binary logistic regression analysis results that there are two predictor variables which is significantly affect the testing result of *Salmonella sp* bacterial contamination on vannamei shrimp (Y), those are the testing result of *Salmonella sp* bacterial contamination on farmers hand swab (X_1) and the subdistrict of vannamei shrimp ponds (X_2). All of the research variables are shown in Table 2.

This bayesian network analysis is perform by using WinBUGS software. The purpose of this analysis is to get all of the parameters estimation in binary logit model ($\hat{g}(x) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2)$), those are β_0, β_1 and β_2, where the parameter β_2 is written as $\beta_2(1)$, $\beta_2(2)$, $\beta_2(3)$ and $\beta_2(4)$ (according to: $X_2(1)$ (Subdistrict *B*), $X_2(2)$ (Subdistrict *C*), $X_2(3)$ (Subdistrict *D*) and $X_2(4)$ (Subdistrict *E*)).

The parameters estimation that is obtained by bayesian network is expected to show the statistical associations between X_1 and X_2 clearly. This bayesian network of binary logistic regression is also use the first reference category as same as the previous binary logistic regression. The DAG of this bayesian network is shown in Fig. 2. The model of DAG is denoted in Fig. 3.

In this bayesian network analysis, three markov chains iteration is used in simulation process. There are two conditions to continue bayesian analysis; those are the posterior distribution of parameters built should be stationary and the parameters should be convergence. Time series plot of history chains is used to check the stationary of posterior distribution. By looking Fig. 4, it obtains that the posterior distribution of parameters are stationary. Figure 5 is the Gelman Rubin statistics of parameters. It shows that the parameters are convergence. Therefore, bayesian network analysis could be continued.

DOI: 10.3844/jmssp.2017.306.311

Table 2: Research variables

Variable	Definition	Category	Scale
Y	The testing result of *Salmonella sp* bacterial contamination on vannamei shrimp	0 : There is no *Salmonella sp* 1 : There is *Salmonella sp*	Nominal
X_1	The testing result of *Salmonella sp* bacterial contamination on farmers hand swab	0 : There is no *Salmonella sp* 1 : There is *Salmonella sp*	Nominal
X_2	The subdistrict of vannamei shrimp ponds	0 : Subdistrict A 1 : Subdistrict B 2 : Subdistrict C 3 : Subdistrict D 4 : Subdistrict E	Nominal

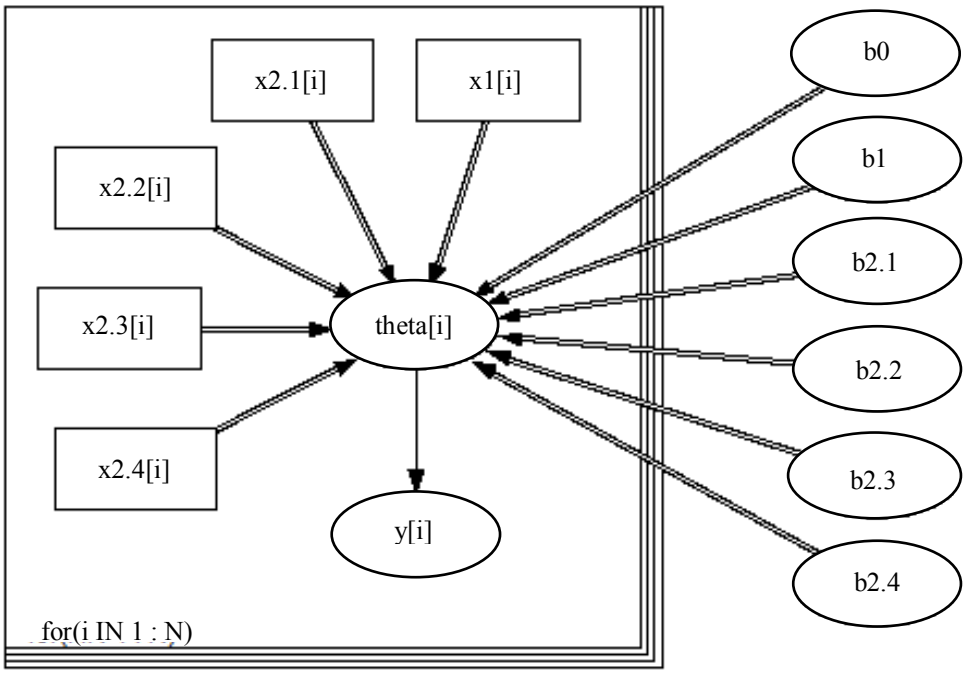

Fig. 2: The DAG for bayesian network of binary logistic regression

```
model;
{
  for( i in 1 : N ) {
    logit(theta[i]) <- b0 + b1 * x1[i] + b2.1 * x2.1[i] + b2.2 * x2.2[i] + b2.3 *
x2.3[i] + b2.4 * x2.4[i]
  }
  for( i in 1 : N ) {
    y[i] ~ dbern(theta[i])
  }
  b0 ~ dnorm( 0.0,1.0E-4)
  b1 ~ dnorm( 0.0,1.0E-4)
  b2.3 ~ dnorm( 0.0,1.0E-4)
  b2.4 ~ dnorm( 0.0,1.0E-4)
  b2.2 ~ dnorm( 0.0,1.0E-4)
  b2.1 ~ dnorm( 0.0,1.0E-4)
    }
```

Fig. 3: The model of DAG

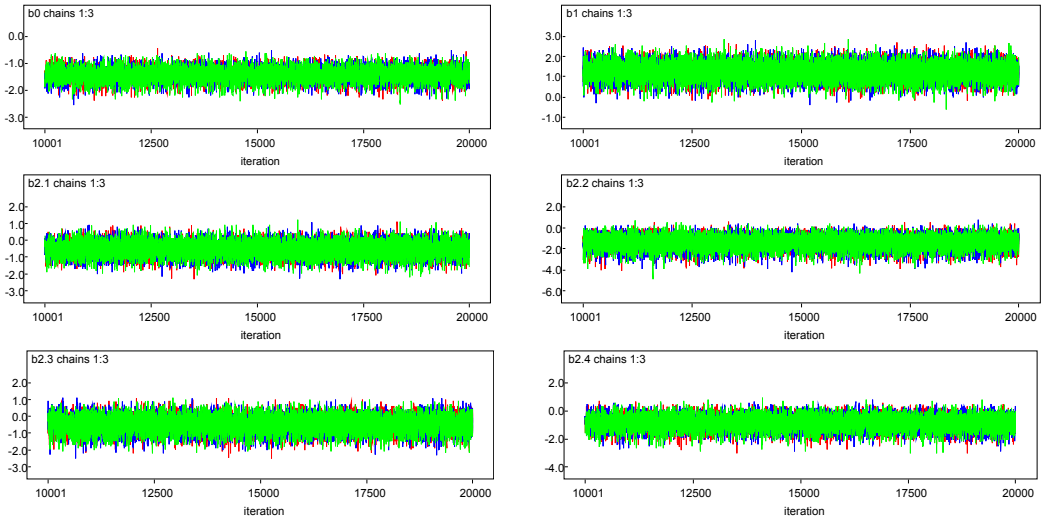

Fig. 4: Time series plot of history chain of parameters

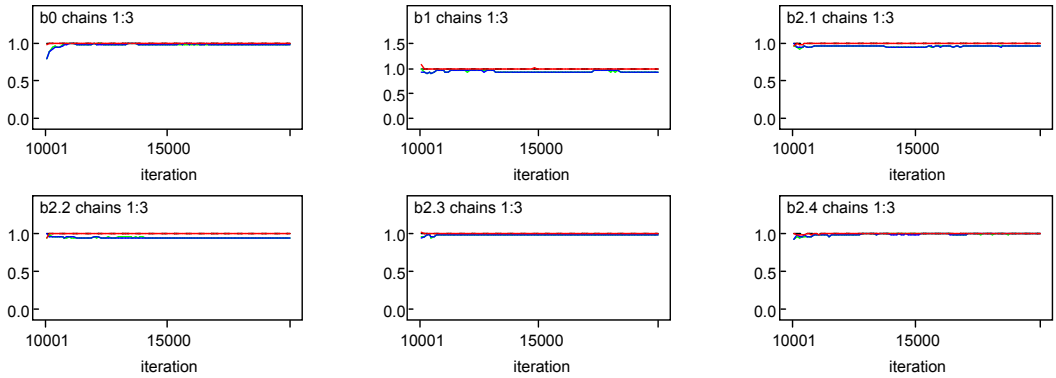

Fig. 5: Gelman rubin statistics of parameters

The results of parameters estimation using bayesian netwok are:

$\beta_0 = -1.428$, $\beta_1 = 1.261$, $\beta_2(1) = -0.514$, $\beta_2(2) = -1.374$, $\beta_2(3) = -0.464$ and $\beta_2(4) = -0.855$

The binary logit model obtained is:

$$\hat{g}(x) = -1.428 + 1.261x_1 - 0.514\,x_2(1) - 1.374\,x_2(2)$$
$$-0.464\,x_2(3) - 0.855\,x_2(4)$$

The probability of presence or absence of *Salmonella sp* in testing result of *Salmonella sp* contamination on vannamei shrimp according to the testing result of *Salmonella sp* bacterial contamination on farmers hand swab and the subdistrict of vannamei shrimp ponds with all of the possible combinationsis shown in Table 3. Those probabilities is calculated by logit model of each combination, which are obtained by regress *Y* and *X* in all possibilities using DAG. The highest probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp (0.350) is obtained if the vannamei shrimp is farmed in Subdistrict A and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*; while the smallest probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp (0.021) is obtained if the vannamei shrimp is farmed in Subdistrict *C* and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp*.

Therefore, the highest probability of the absence of *Salmonella sp* in testing result on vannamei shrimp (0.979) is obtained if the vannamei shrimp is farmed in Subdistrict *C* and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp*; while the smallest probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp (0.650) is obtained if the vannamei shrimp is farmed in Subdistrict *A* and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*.

Table 3: The probability of presence or absence of *Salmonella sp* in testing result of *Salmonella sp* contamination on vannamei shrip

The testing result of *Salmonella sp* contamination on farmer hand swab	The subdistrict of vannamei shrimp ponds	The probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp	The probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp
There is no *Salmonella sp*	Subdistrict A	0.864	0.136
	Subdistrict B	0.874	0.126
	Subdistrict C	0.979	0.021
	Subdistrict D	0.874	0.126
	Subdistrict E	0.977	0.023
There is *Salmonella sp*	Subdistrict A	0.650	0.350
	Subdistrict B	0.671	0.329
	Subdistrict C	0.932	0.068
	Subdistrict D	0.671	0.329
	Subdistrict E	0.924	0.076

Conclusion

The result of bayesian network analysis of binary logistic regression obtain the statistical associations between the significant predictor variables and the contamination of *Salmonella sp* bacterial on vannamei shrimp which is show in probability as following:

- The probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict A and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*, is 0.350
- The probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict B and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*, is 0.329
- The probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict C and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*, is 0.068
- The probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict D and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*, is 0.329
- The probability of the **presence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict E and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is *Salmonella sp*, is 0.076
- The probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict A and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp*, is 0.136
- The probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict B and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp*, is 0.126
- The probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict C and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp* is 0.021
- The probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict D and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp*, is 0.126
- The probability of the **absence** of *Salmonella sp* in testing result on vannamei shrimp obtained if the vannamei shrimp is farmed in Subdistrict E and the testing result of *Salmonella sp* contamination on farmer hand swab show that there is no *Salmonella sp*, is 0.023

Acknowledgment

We are grateful to LPPM ITS who gives the chance to finish this research in part of the research for beginner (Penelitian Pemula) in Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

Author's Contributions

Pratnya Paramitha Oktaviana: Participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Kartika Fithriasari: Guide the first author and contributed to the writing of the manuscript.

Ethics

The authors confirm that this paper is original and approved the manuscript. There is no ethical issues involved.

References

Chen, C., G. Zhang, R. Tarefder, J. Ma and H. Wei *et al*., 2015. A multinomial logit model-bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Elsevier: Accident Anal. Prevent., 80: 76-88. DOI: 10.1016/j.aap.2015.03.036

Hosmer, D.W. and D.S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn, John Wiley and Sons, Inc., New York, ISBN-10: 0471356328, pp: 373.

Liu, Y., 2012. Doodlebugs application. Department of Statistics, University of Missouri, Columbia.

Stephenson, T.A., 2000. An introduction to bayesian betwork theory and usage. Research Report, Idiap Research Institute, Martigny-Valais, Switzerland.