The Estimation Process in Bayesian Structural Equation Modeling Approach

Ferra Yanuar

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Andalas University, Kampus Limau Manis, 25163, Padang, Indonesia.

Email: ferrayanuar@yahoo.co.id

Abstract. Structural equation modelling (SEM) is a multivariate method that incorporates ideas from regression, path-analysis and factor analysis. A Bayesian approach to SEM may enable models that reflect hypotheses based on complex theory. The development and application of Bayesian approaches to SEM has, however, been relatively slow but with modern technology and the Gibbs sampler, is now possible. The Gibbs sampler can be used to obtain samples of arbitrary size from the posterior distribution over the parameters of a structural equation model (SEM) given covariance data and a prior distribution over the parameters. Point estimates, standard deviations and interval estimates for the parameters can be computed from these samples. This study shows that the conditional distributions required in the Gibbs sampler are familiar distributions, hence the algorithm is very efficient. A goodness of fit statistic for assessing the proposed model is presented. An illustrative example with a real data is presented.
Key words: Structural equation modeling, Bayesian approach, Gibbs sampler, prior distribution.

1. Introduction

Structural equation modeling (SEM) is a collection of statistical techniques that provides a powerful set of tools for researchers in education, social, behavioral and other disciplines. SEM can suitably be used for the analysis of complex phenomena which involves hypothesized relationships between one or more independent latent variables and one or more dependent latent variables [1]. The general goal of SEM is to test the hypothesis that the observed variance matrix for a set of measured variables is equal to the covariance matrix implied by a hypothesized model. This relationship can be formally stated as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \qquad (1)$$

where $\boldsymbol{\Sigma}$ represents the population covariance matrix of a set of observed variables and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represents the population covariance matrix implied by θ, a vector of model parameters. The vector $\boldsymbol{\theta}$ thus defines the form of a particular SEM through the specification of means and intercepts, variances and covariances, regression parameters, and factor loadings.

Certain assumptions need to be fulfilled before SEM can be applied to a particular data set. Under the traditional application of SEM, the data set are assumed drawn from a parent population which is assumed normally distributed; hence, the joint distribution of the variables which represent the data follows a multivariate normal distribution. If that condition is fulfilled, the standard estimation method in SEM which is recommended to be used to estimate parameters and standard errors is maximum likelihood (ML) method.

It is often found that the data gathered in a survey do not follow the assumption of multivariate normality. In the social science research, for example, social and behavioral attitudes are usually measured using ordered categories or dichotomous scaled. Clearly, such data are, by definition, not normally distributed

[2]. Often, the measured variables are continuous but their distributions depart dramatically from normality. We cannot use ML estimation method anymore for such conditions because model fit indices, parameter estimates and standard errors tend to be bias as non normality increase [3]. Hancock & Mueller [4] found that the effects of violating the assumption of non normality include large chi-square values (so too many models are rejected) and standard errors are too small (so significance testing of path coefficients will result an increase in Type I error rates).

The basic objective of this study is to demostrate the Bayesian approach for analyzing SEM, then the methodology is applied to a real data set. Different from classical SEM where computational algorithm is developed based on sampel covariance matrix, in Bayesian SEM, we focus on the use of the raw observations rather than the sampel covariance matrix [5] and apply some powerful tools in statistical computing. We treat the latent variables in the model and the latent measurements as missing data then we analyze the model on the basis of the complete data set. We use MCMC techniques to estimate the unknown parameters in the model. Gibbs sampler as a method in MCMC is applied to obtain a sequence of random samples for summarizing the posterior distribution of parameter model. Moreover, Bayesian technique has flexibility to use useful prior information for achieving better results.

2. Bayesian SEM Approach

In the basic SEM model, it consists of measurement and structural equation. Measurement equation in SEM approach is given by:

$$\boldsymbol{x_i} = \boldsymbol{\Lambda \omega_i} + \boldsymbol{\varepsilon_i}, \quad i = 1, \dots, n \tag{2}$$

where $\boldsymbol{x_i}$ is an $p \times 1$ vector of indicators describing the $q \times 1$ random vector of latent variables $\boldsymbol{\omega_i}$, $\boldsymbol{\Lambda}$ is $p \times q$ matrices of the loading coefficients as obtained from the regressions of $\boldsymbol{x_i}$ on $\boldsymbol{\omega_i}$ and $\boldsymbol{\varepsilon_i}$ is $p \times 1$ random vectors of the measurement errors which follow $N(0, \boldsymbol{\psi_\varepsilon})$. It is assumed that for $i = 1, \dots, n$, $\boldsymbol{\omega_i}$ is independent follows a normal distribution $N(0, \boldsymbol{\Phi})$ and uncorrelated with the random vector $\boldsymbol{\varepsilon_i}$.

Let the latent variable $\boldsymbol{\omega_i}$ be partitioned into $(\boldsymbol{\eta_i}, \boldsymbol{\xi_i})$ where $\boldsymbol{\eta_i}$ and $\boldsymbol{\xi_i}$ are $m \times 1$ and $n \times 1$ vectors of latent variables respectively. The structural equation of SEM which explaining the interrelationship among the latent factors is expresessed by:

$$\boldsymbol{\eta_i} = \boldsymbol{B\eta_i} + \boldsymbol{\Gamma \xi_i} + \boldsymbol{\delta}_\mathrm{i}, i = 1, \dots, n. \tag{3}$$

where $\boldsymbol{B}$ is $m \times m$ matrix of structural parameters governing the relationship among the endogenous latent variables which is assumed to have zeros in the diagonal, $\boldsymbol{\Gamma}$ is $m \times n$ regression parameter matrix for relating the endogenous latent variables and exogenous latent variables, and $\boldsymbol{\delta}_\mathrm{i}$ is $m \times 1$ vector of disturbances which is assumed $N(0, \boldsymbol{\psi_\delta})$ where $\boldsymbol{\psi_\delta}$ is a diagonal covariance matrix**.** It is also assumed that $\boldsymbol{\delta}_\mathrm{i}$ is uncorrelated with $\boldsymbol{\xi_i}$**.** Since only one endogenous latent variable involve in this study, or $\boldsymbol{B\eta_i} = 0$, so equation (3) can be rewrite become $\boldsymbol{\eta_i} = \boldsymbol{\Gamma \xi_i} + \boldsymbol{\delta}_\mathrm{i}$.

Under Bayesian approach in SEM, we consider $\boldsymbol{Y} = (\boldsymbol{y_1}, \boldsymbol{y_2}, \dots \boldsymbol{y_n})$ and $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots \boldsymbol{x_n})$ be data matrices and let $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_n)$ be the matrix of latent variables and the structural parameter $\boldsymbol{\theta}$**,** a vector that includes all the unknown parameters in $\boldsymbol{\Phi}$, $\boldsymbol{\psi_\delta}$, $\boldsymbol{\psi_\varepsilon}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_\omega$ [5, 6]. We apply Markov Chain Monte Carlo (MCMC) methods to obtain the Bayesian estimates $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$. To achieve our goal, a sequence of random observations from the joint posterior distribution $[\boldsymbol{\theta}, \Omega | \boldsymbol{Y}]$ will be generated via the Gibbs sampler. The Gibbs sampler is a MCMC technique that generates a sequence of random observations from the full conditional posterior distribution of unknown model parameters. The user can create and implement the algorithm easily using WINBUGS [7]. The Gibbs sampler process starts with

the setting of initial starting values $\left(\boldsymbol{\theta}^{(\mathbf{0})}, \boldsymbol{\Omega}^{(\mathbf{0})}, \boldsymbol{Y}^{(\mathbf{0})}\right)$, and then conduct the simulation for $\left(\boldsymbol{\theta}^{(\mathbf{1})}, \boldsymbol{\Omega}^{(\mathbf{1})}, \boldsymbol{Y}^{(\mathbf{1})}\right)$. At the *r*th iteration, by making use of the current values $\left(\boldsymbol{\theta}^{(r)}, \boldsymbol{\Omega}^{(r)}, \boldsymbol{Y}^{(r)}\right)$, the Gibbs sampler is carried out as follows:

a. Generate a random variate $\Omega^{(r+1)}$ from the conditional distribution $\left(\Omega|, \theta^{(r)}, Y^{(r)}\right)$
b. Generate a random variate $\theta^{(r+1)}$ from the conditional distribution $\left(\theta|\Omega^{(r+1)}, Y^{(r)}\right)$
c. Generate $\left(\boldsymbol{Y}^{(\boldsymbol{r+1})}\right)$ from $p\left(Y|\boldsymbol{\Omega}^{(\boldsymbol{r+1})}, \boldsymbol{\theta}^{(\boldsymbol{r+1})}\right)$

Under mild regularity conditions, the samples converge to the desired posterior distribution. The derivation of the conditional distribution that is required in the Gibbs sampler process above is discussed in Lee & Shi [8] or Lee [5]. In the process when determining the posterior distribution, the selection of prior distribution for $(\boldsymbol{\Lambda}, \boldsymbol{\psi_\varepsilon})$ and $\boldsymbol{\Phi}$ have to be made. In this study, we take the prior distribution for those three parameters via the following conjugate type distribution. Letting $\psi_{\varepsilon k}$ be the *k*th diagonal element of $\boldsymbol{\psi_\varepsilon}$ and $\boldsymbol{\Lambda}_k$ be the *k*th row of $\boldsymbol{\Lambda}$, we consider:

$$\psi_{\varepsilon k}^{-1} \sim Gamma(\alpha_{0\varepsilon k}, \beta_{0\varepsilon k}) \tag{4}$$

$$\left(\Lambda_k|\psi_{\varepsilon k}^{-1}\right) \sim N\left(\Lambda_{0k}, \psi_{\varepsilon k}\mathbf{H}_{0yk}\right) \tag{5}$$

$$\boldsymbol{\Phi}^{-1} \sim W_q(R_0, \rho_0) \tag{6}$$

where $Gamma(,)$ is the gamma distribution, $W_q(,)$ is an q dimensional Wishart distribution, parameters $\alpha_{0\varepsilon k}, \beta_{0\varepsilon k}, \Lambda_{0k}, \rho_0$, positive definite matrix $\mathrm{H_{0yk}}$ and R_0 are hyperparameters which are assumed to be described by an uninformative prior distribution.

The next process in Bayesian SEM is convergence test of the model parameters. The convergence is assessed using a variety of diagnostics as detailed in the CODA package, plotting the time series to assess the quality of the individual parameters with different starting values graphically, and provide a diagnosis based on the trace plots [9, 10, 11]. We also use the Brooks Gelman-Rubin (BGR) convergence statistics [12]. This convergence statistics test compares the variation between and within multiple chains, denoted by $\boldsymbol{R}$. The estimated parameters converge if the value of $\boldsymbol{R}$ is close to 1. In addition, the accuracy of the posterior estimates are inspected by assuring that the Monte Carlo error (an estimate of the difference between the mean of the sampled values and the true posterior mean) for all the parameters to be less than 5% of the sample standard deviation [6].

For assessing the plausibility of our proposed model which includes the measurement equation and structural equation, we plot the residual estimates versus latent variable estimates to give information for the fit of the model. The residuals estimates for measurement equation $(\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}})$ can be obtained from

$$\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\xi}}_{\boldsymbol{i}}, \quad i = 1, \ldots, n \tag{7}$$

where $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\xi}}_{\boldsymbol{i}}$ are Bayesian estimates obtained via the MCMC methods. The hypothesized models provide a good fit if the plots are centered at zero and lie within two parallel horizontal lines. The estimates of residuals in the structural equation $\left(\hat{\boldsymbol{\delta}}_{\boldsymbol{i}}\right)$ can be obtained from following equation:

$$\hat{\boldsymbol{\delta}}_{\boldsymbol{i}} = (\boldsymbol{I} - \hat{\boldsymbol{B}})\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\xi}}_i, \quad i = 1,\ldots,n \tag{8}$$

where $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\eta}}_i$, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\xi}}_i$ are Bayesian estimates that are obtained from the corresponding simulated observations through the MCMC. The proposed model fitted the data well or provided a reasonably good

fit if the plots lie within two parallel horizontal lines that are centered at zero and no trends are detected [6].

3. Example: Modeling of Health Index

In this section we will illustrate the application of Bayesian SEM for constructing the model of health index. The data set used for the analysis is based on the Third National Morbidity Survey (NHMS III) that took place in Malaysia and conducted at the year 2006. NHMS III which organized by Institute for Public Health, Ministry of Health is a population based cross-sectional study using two stage stratified sampling design proportionate to population size throughout all states in Malaysia. The details of the methodology of survey have been reported previously [13]. The data which used to be analyzed are focused on the respondents who are living in Hulu Langat only, a district in Malaysia. There are 530 respondents involved in the analysis those who are 18 years old and older and had given a complete information required in this study. The information gathered in the survey includes information related to socio-demographic status, lifestyle, mental health condition and biomarkers of individuals. A questionnaire-form was used to obtain the information from the respondents. In addition, medical screening for measuring body height and weight, blood pressure, cholesterol level, HDL (high density lipoprotein) and blood glucose was done either during the house visit or during the consultation at the clinic. Respondents were also asked about the number of health problems experienced and perception about their health condition at the time of the interview.

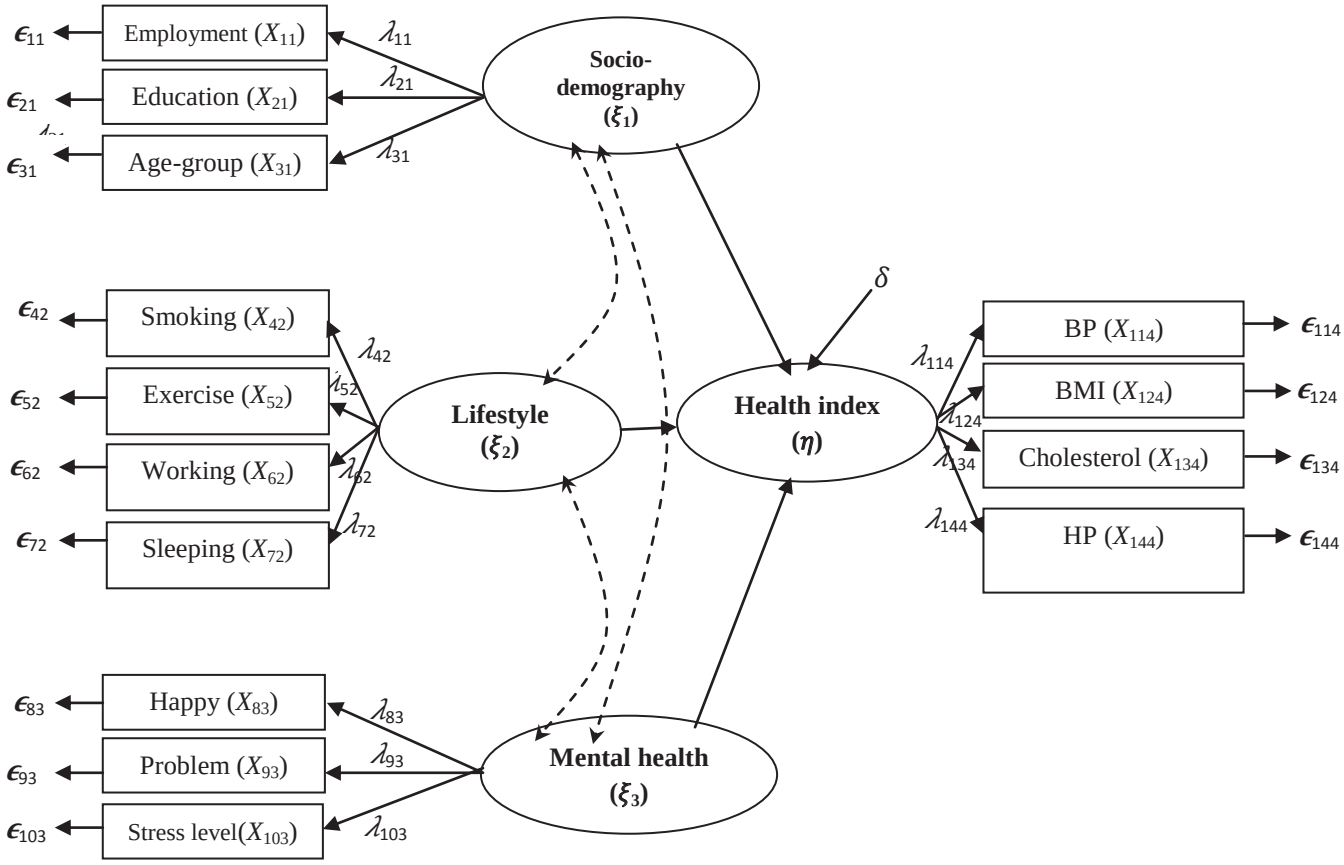
3.1. Factors related to health

There are many factors which are related to health, some are observable and some are not, such as level of blood pressure, level of cholesterol, etc. Biomarkers can be considered as an example of observable measures which can be used to describe the level of health of an individual. The factors that are not directly observed such as lifestyle, socio-demography and mental health condition could be measured through the indicator variables. Following is a brief explanation for each unobservable variable considered in this study and its respective indicators.

Socio-demography status is an important determinant of health index. In several studies, such as Broadman [14], Cheadle et al. [15], Uitenbroek et al. [16] and Shi [17], it is found that people who are poorer, less well educated or having a lower job status tend to have a lower health index. The indicators of interest considered in this study for assessing the socio-demography factor are education level, employment status and age-group. It is quite well informed of the important role of lifestyle as one of a contributary factor in influencing the level of health of an individual. Unhealthy lifestyle, such as being a smoker or rarely having physical exercise, could have a negative influence to the health condition of an individual [15, 16, 17,18]. The indicators of lifestyle, based on the list of health related behaviors as suggested by Nakayama et al. [18] which has also been used by Broadman [14] that are considered in this study, are physical exercise, smoking habits, average working hours per day and average sleeping hours per day.

Apart from lifestyle, mental health is often recognized as one of the major health determinants. A study by Nakayama et al. [18] showed that health was influenced by many factors, including mental health. Hays et al. [19] found that there was a significant correlation between physical health and mental health. Hence, it is reasonable to assume that mental health as a contributory factor for determining the health index. Nakayama et al.[18] and Broadman [14] have suggested the use of stress level and the number of experiences on serious problem as the indicator for mental health. They found that people who have a high stress level as well as have experienced problems possess a low level of mental health.

In this study, it is hypothesized that socio-demography, lifestyle and mental health are latent factors that are related to the health index of an individual. The health index could also be measured directly based on certain indicators such as body mass index (BMI), blood pressure, cholesterol level, the number of common health problems experienced by the respondent. These hypothesis model is ilustrated in Figure 1.



BP = Blood pressure
BMI = Body mass index
HP = Health problem

Figure 1. A Diagrammatic Illustration for Health Index Model.

3.2. Results

The first step we do test the sensitivity of the Bayesian anaysis by considering three types of prior inputs. We found that the statistics yielded based on Bayesian SEM is not sensitive to the different prior inputs or we could say that Bayesian SEM applied in this study is robust enough. Then, test of convergence

statistics for all parameters of interest shows that the vaue of R are close to 1. Plots of sequences of observations corresponding to some parameters generated by two different initial values are also indicate converge. Then, test of accuracy for the posterior estimates prove that Monte Carlo error for all parameters are less than 5% of the sample standard deviation. Based on all tests we observe that Bayesian SEM estimates are close to the true values and that the standard error estimates are reasonable and we conclude here that the proposed measurement equation and structural equation are adequate. Figure 1 provides the fitted model for the structural model in Bayesian SEM which includes the standardized parameter estimates and their standard errors, covariance among latent variables and structural errors of the model. Table 1 shows the estimate values of loading factors which relate indicator variables and corresponding latent variables.

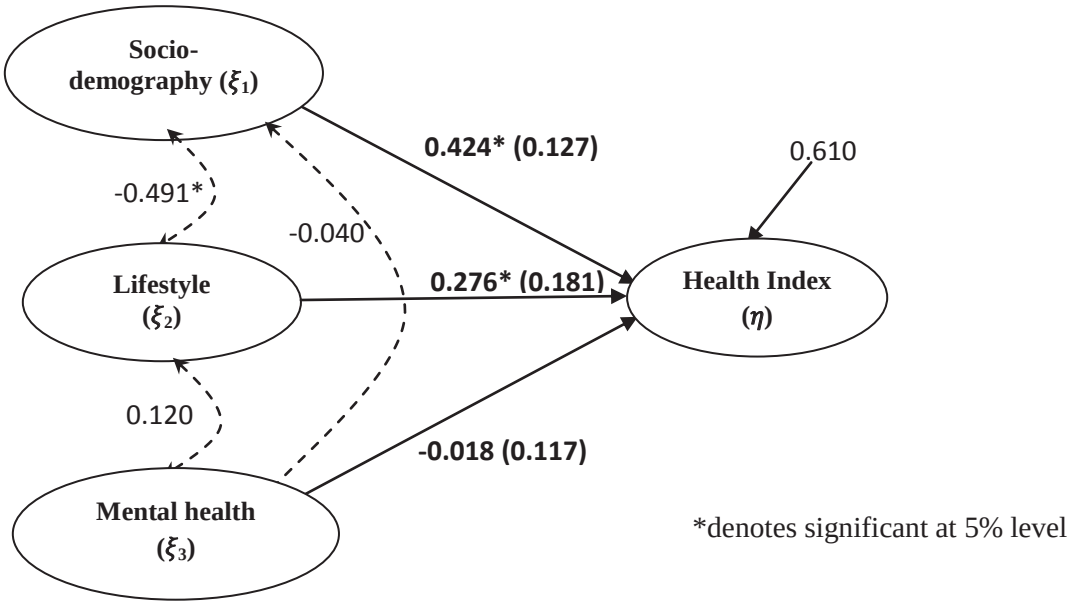


Figure 2. Parameter Estimates for Structural Model

Based on Figure 2, we could also present the structural model that address the relationship between socio-demography status, lifestyle, mental health with health index in the following equation :

$$\widehat{\boldsymbol{\eta}}_{Bay\ SEM} = 0.424\xi_1 + 0.276\xi_2 - 0.018\xi_3 \quad (13)$$

This equation indicate that the greatest effect to the health index is socio-demography then followed by lifestyle and mental health. Socio-demography and ifestyle give the significant effect to the health index meanwhile mental health does not. It is possible since there are other indicator variables for mental health should be incorporated into the model. This study aso inform that there is significant correlation between socio-demographic status and lifestyle.

Table 1 shows the value of coefficient of factor loading and the associate standard errors for each indicator variable in the measurement equations obtained based on Bayesian SEM approach.

Table 1. Factor Loading Estimates in Measurement Model

Construct	Estimate	
	Factor Loadings (SE)	**95% CI**
Socio–demography → Employment	1	-
Socio–demography → Education	0.424 (0.133)*	(0.650 , 0.987)
Socio–demography → Age group	-0.334 (0.123)*	(-0.582 , -0.103)
Lifestyle → Smoking	1	-
Lifestyle → Exercise	-0.071 (0.128)	(-0.430 , 0.306)
Lifestyle → Working hours	0.358 (0.107)*	(0.155 , 0.771)
Lifestyle → Sleeping hours	0.309 (0.195)	(-0.064 , 0.606)
Mental Health → Happy	1	-
Mental Health → Problem	0.884 (0.157)*	(0.590 , 0,921)
Mental Health →Stress leve	0.824 (0.167)*	(0.500 , 1.029)
Health Index → BP	1	-
Health Index → BMI	0.862 (0.103)*	(0.667 , 0.989)
Health Index → Cholesterol	0.401 (0.108)*	(0.189 , 0.601)
Health Index → Health Problem	0.662 (0.106)*	(0.456 , 0.869)

SE =standard error, CI=confidence interval, BP= blood pressure, BMI= body mass index,

*Significant at 5% level

4. Discussion

In classical SEM, computational algorithm is developed based on sampel covariance matrix and normal assumptions for the observations. But in many studies, It is often found that the data gathered in a survey do not follow the assumption of multivariate normality. Bayesian appproach in SEM is believed as a potential tools to overcome nonnormal assumption [8, 9, 11].

The basic objective of this present study is to demostrate the Bayesian approach for analyzing SEM. In contrast to maximum likelihood method, in Bayesian estimations, parameters are considered as random with prior distribution and a prior density function [5]. Once the data is collected, it is combined with prior distribution using Bayes theorem, next posterior distribution is calculated reflecting the prior knowledge and empirical data. Joint posterior distribution is summarized using MCMC simulation techniques in terms of lower dimensional summary statistics as posterior mean and posterior standard deviations.

The methodology of Bayesian SEM then applied to a real data set. We observe that the structural and measurement equation obtained from this study are adequate and in general we could accept the proposed model.

Acknowledgments
The authors thank to the Institute for Public Health, Ministry of Health, Malaysia, who furnished us the health survey data used in this study. We also thank several anonymous referees for their constructive comments which have improved the final version of this paper.

References

[1] Ullman, J. B. 2006. Structural equation modeling: reviewing the basics and moving forward. *Journal of Personality Assessment*, 87, 35-50.

[2] Kaplan, D. 2000. *Structural equation modeling*. Thousand Oaks, CA: Sage.

[3] U.H. Olsson , T. Foss, S.V. Troye, and R.D. Howell. *The performance of ML, GLS, and WLS estimation in SEM under conditions of misspecification and non-normality*. Structural Equation Modeling 7 (2000), pp. 557–595.

[4] Hancock G.R., & Mueller R.O. 2006. *Structural equation modeling: A second course*. Information Age Publishing, Inc.

[5] Lee. 2007. *Structural equation modeling: A Bayesian approach*. John Wiley & Sons, Ltd, New York.

[6] Yanuar, F., Ibrahim, K. & Abdul, A.J. 2013. Bayesian structural equation modeling for the health index. *Journal of Applied Statistics*, 40(6): 1254-1269.

[7] Spiegelhalter, Thomas, Best, and Lunn. *WinBUGS User Manual*, *Version* 1.4, software available at http://www.mrc-bsu.cam.ac.uk/bugs, 2003.

[8] S.Y. Lee and J.Q. Shi. *Bayesian analysis of structural equation model with fixed covariates*. Structural Equation Modeling. 7 (2000), pp. 411–430.

[9] A. Ansari, K. Jedidi, and L. Dube L. *Heterogeneous factor analysis models: A Bayesian approach*. Psychometrika. 67 (2002), pp. 49–78.

[10] R.E. Kass and A.E. Raftery. *Bayes factors*. Journal of the American Statistical Association. 90 (1995), pp. 773–795.

[11] R. Scheines, H. Hoijtink, and A. Boomsma. *Bayesian estimation and testing of structural equation models*. Psychometrika 64 (1999), pp. 37–52.

[12] J. Palomo, D.B. Dunson, and K. Bollen. *Bayesian structural equation modeling*. Handbook of Computing and Statistics with Application. 1 (2007), pp. 163–188.

[13] Institute for Public Health, 2008, The Third National Health and Morbidity Survey (NHMS III) 2006, Vol 1. Ministry of Health, Malaysia.

[14] J.D. Boardman. *Stress and physical health: the role of neighborhoods as mediating and moderating mechanisms*. Social Science & Medicine. 58 (2004), pp. 2473–2483.

[15] A. Cheadle, D. Pearson, E. Wagner, B.M. Psaty, P. Diehr, and T. Koepsell. *Relationship between socioeconomic status, health status, and lifestyle practices of American Indians: Evidence from a plains reservation population*. Public Health Report. 109 (1994), pp. 405–413.

[16] D.G. Uitenbroek, A. Kerekovska, and N. Festchieva. 1996. Health lifestyle behaviour and socio-demography characteristics. A study of Varna, Glasgow and Edinburgh. *Social Science & Medicine*. 43, pp. 367–377.

[17] Shi. 1998. Socio-demography characteristics and individual health behaviors. *Socio demographic factors and health behaviors* 9, pp. 933–942.

[18] Nakayama K., Yamaguchi K., Maruyama S. & Morimoto K. 2001. The relationship of lifestyle factors, personal character, and mental health status of employees of a major Japanese electrical manufacturer. *Environmental Health and Preventive Medicine* , 5: 144-149.

[19] Hays, R.D., Revicki, D., & Coyne, K.S. 2005. Application of structural equation modeling to health outcomes research. *Eval Health Prof* , 28; 295-309.