Bayesian structural equation modeling for the health index

Ferra Yanuara∗, Kamarulzaman Ibrahimb and Abdul Aziz Jemainb

aDepartment of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Andalas, Kampus Limau Manis, 25163 Padang, Indonesia; bSchool of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

(Received 29 September 2010; accepted 11 March 2013)

There are many factors which could influence the level of health of an individual. These factors are interactive and their overall effects on health are usually measured by an index which is called as health index. The health index could also be used as an indicator to describe the health level of a community. Since the health index is important, many research have been done to study its determinant. The main purpose of this study is to model the health index of an individual based on classical structural equation modeling (SEM) and Bayesian SEM. For estimation of the parameters in the measurement and structural equation models, the classical SEM applies the robust-weighted least-square approach, while the Bayesian SEM implements the Gibbs sampler algorithm. The Bayesian SEM approach allows the user to use the prior information for updating the current information on the parameter. Both methods are applied to the data gathered from a survey conducted in Hulu Langat, a district in Malaysia. Based on the classical and the Bayesian SEM, it is found that demographic status and lifestyle are significantly related to the health index. However, mental health has no significant relation to the health index.

Keywords: health index; structural equation modeling; Bayesian SEM; Gibbs sampler; prior information

1. Introduction

The information on the health status of an individual is often gathered based on a health survey. It can be used by decision-makers to allocate resources prudently when planning of activities which aims at improving the overall health status of a particular community is made. For ease of interpretation, this information can be summarized in a single value called a health index. Therefore, it is important to identify the factors that could affect the health index. Various studies indicate that many factors influence the health index of an individual, including socio-demography, lifestyle and mental health. Since these influential factors are interrelated and latent because they cannot be measured directly, it is quite complicated to determine the health index. Consequently, an appropriate technique which allows for this interrelationship known as the structural equation

∗Corresponding author. Email: ferrayanuar@yahoo.co.id

© 2013 Taylor & Francis
modeling (SEM) has to be applied when estimating the health index. SEM is the statistical tool which is suitable to be applied for summarizing the health index of an individual. In this study, the classical SEM is used to construct the model for describing the health index. SEM has been used by many researchers to analyze a complex phenomenon which involves hypothesized relationships between independent latent variables and dependent latent variables [5,13,35]. Since the general goal of SEM is to test the hypothesis that the observed covariance matrix for a set of measured variables is equal to the covariance matrix defined by the hypothesized model, computational algorithm in SEM is developed on the basis of the sample covariance matrix $S$. SEM uses the assumption that the observations are independent and identically distributed according to the multivariate normal distribution [18]. If this assumption is not fulfilled, the sample covariance matrix cannot be determined in the usual way and difficult to be obtained [28]. Therefore, many researchers such as Scheines et al. [31], Lee and Shi [19], Ansari et al. [2,3] and Lee et al. [21] proposed the use of Bayesian approach in SEM to overcome these problems. This involves the use of Gibbs sampler [12] to obtain samples of arbitrary sizes for summarizing the posterior distribution for describing the parameters of interest. From these samples, the user can compute the point estimates, standard deviations and interval estimates for the purpose of making an inference. The Bayesian approach is attractive since it allows the user to use the prior information for updating the current information regarding the parameters of interest.

The main purpose of this study is to illustrate the value of the classical SEM and the Bayesian SEM for developing a model which describes the health index of an individual living in Hulu Langat, Malaysia. The interrelationship among the latent variables such as mental health, socio-demographic characteristics and lifestyle, and between the latent variables and their respective manifest variables are determined using the data found from the survey that were undertaken in Hulu Langat.

2. Data and instrument

The analysis is applied on a data set obtained from the Health Risk Assessment (HRA) survey that took place in Hulu Langat, a district in Malaysia [9]. The HRA survey was organized by the Department of Community Health, Medical Faculty, Universiti Kebangsaan Malaysia (UKM) in collaboration with Hospital Universiti Kebangsaan Malaysia (HUKM) [9] in the year 2001. HUKM was involved as the main center that coordinated the survey.

The objective of the survey was to promote disease prevention activities and healthy living for people in the area of Hulu Langat. The data obtained in the HRA survey have also been applied by Yanuar et al. [37] for constructing the health index. It has to be clarified here that the types of data involved in their analysis are continuous, binary and ordinal. In this study, however, all the data considered are of the ordinal type and details on them are given later in this section. Considering only one type of data makes the analysis slightly easier. In addition, both the classical and the Bayesian SEM are applied on the same type of data, making the results found based on the two approaches comparable. Based on the survey, it is reported that about 5.7% of the people living in Hulu Langat experienced having three most common diseases which include gastritis, diabetes and asthma. The data that were provided by 5035 respondents who had given a complete information required in the survey are considered in the analysis. The respondents eligible to participate in the survey are those who are 18 years old and older, representing the adult population of Hulu Langat. The respondents are chosen randomly based on the house visit or consultation in HUKM between April 2000 and August 2001. The choice of house visit is done randomly based on the list of household that were available with the enumerators.

The information gathered in the survey includes information about lifestyle, socio-demographic status, mental health condition and biomarkers of individuals living in Hulu Langat. The indicators used for describing the socio-demographic factor are employment status, education level and age
Respondents are asked about their employment status, and the responses are indicated by 1, 2 and 3 for ‘unemployed’, ‘ordinary’ and ‘professional’, respectively. With respect to the education level, the responses obtained are coded as 1 as ‘never attend school or attend elementary school’, 2 as ‘attend high school’ and 3 as ‘attend college/university’. The responses to employment status are coded as 1, 2 and 3 for ‘unemployed’, ‘ordinary’ and ‘professional’, respectively. The age of the respondents are classified into three groups which are between 18 and 31 years old, between 31 and 56 and more than 56 years old, coded as 1, 2 and 3, respectively.

The indicators for lifestyle as hypothesized in this study are based on the list of health-related behaviors that have been proposed by several authors. The list of health-related behaviors that has been suggested by Nakayama et al. [25] is the list that has been used by Boardman [4] and Shi [32], which includes physical exercise, smoking habits, average sleeping hours and average working hours in a day, is considered in this study. The respondents were asked about their smoking habits and the responses are coded as 1, 2 and 3 for denoting ‘smoker’, ‘quit’ and ‘non-smoker’, respectively. Regarding frequency of physical exercise, respondents were asked ‘how many times a week on the average do you have physical exercise?’ The responses for this question consist of three categories that are coded as 1, 2 and 3 to indicate ‘none’, ‘1 or 2 times in a week’ and ‘more than 3 times in a week’, respectively. Working hours in a day are coded into 1 as ‘more than 14 h’, 2 as ‘9–14 h’ and 3 as ‘less than 9 h’. Average sleeping hours in a day are grouped as well, with 1 referring to ‘less than 7 h’, 2 as ‘more than 8 h’ and 3 as ‘7–8 h’.

Nakayama et al. [25] and Boardman [4] have proposed that level of stress, happiness in life and the number of serious problems that were faced during the last year could be used as the indicators for mental health. The stress level is divided into three categories which are coded as 1, 2 and 3, referring to ‘high stress level’, ‘middle’ and ‘normal’, respectively. Happiness in life is coded into 1 as ‘not happy’, 2 as ‘average’ and 3 as ‘happy’. Meanwhile, the number of experiences of serious problems during the last year has been categorized into three categories as well and coded as 1, 2 and 3, referring to ‘having more than 2 problems’, ‘having 1 or 2 problems’ and ‘no problem’, respectively. In addition, medical screening was carried out to measure weight and height, cholesterol and blood pressure level during consultation at the clinic in the hospital or during a house visit. The number of health problems that was faced by the respondents are also asked.

The health index which is assumed to be related to mental health, socio-demography and lifestyle could also be measured directly based on certain indicators. In this study, the indicators of the health index are blood pressure, cholesterol level, body mass index (BMI) and the number of common health problems that the respondent ever had. BMI is defined as the body mass of an individual divided by the square of his or her height. The values of BMI that are coded as 1, 2 and 3, which denote ‘obese’, ‘risky (underweight or overweight)’ and ‘normal’, respectively, are based on the guidelines provided by the Centers for Disease Control and Prevention [7]. The classifications that are made for blood pressure level are ‘high blood pressure’ which is coded as 1, ‘pre-hypertension’ which is coded as 2 and ‘normal’ which is coded as 3, based on the categories proposed by the American Heart Association [1]. The classifications that are used for the total cholesterol level consist of three ordered categories as well and decided using the categories suggested by the National Institutes of Health [26]. Regarding the information on the health level, the respondents are asked whether they have ever had one or more of the 14 common health problems considered in the questionnaires of the survey. The respondents who have more than four health problems are categorized as ‘unhealthy’ and coded as 1, who have two to four health problems are considered as ‘less healthy’ and coded as 2, and those with fewer than two health problems are considered as ‘healthy’ and coded as 3.

In Figure 1, the hypothesized model which involves measurement and structural components is used to illustrate the health index model. Since it is also reasonable to hypothesize that lifestyle, socio-demographic status and mental health are correlated, we suggest that the presence of the
interrelationships among these three latent variables to be indicated as in Figure 1 by the dash line with double-headed arrows connecting the latent variables.

3. Statistical analysis

3.1 Classical SEM

Before SEM method is applied, a path model is delineated using the hypothesized relationships which involve the interrelationships between the latent variables based on some previous studies, as presented in Figure 1. The latent variables considered are lifestyle, socio-demography, mental health and health index. In this study, we apply the confirmatory factor analysis (CFA) in SEM that allows the identified latent constructs to be correlated and any parameters to be fixed at a preassigned value. It is more pertinent to apply the CFA rather than exploratory factor analysis since we have identified the latent factors, as argued earlier, which we believe to covary with each other.

In this section, the basic SEM model with ordered categorical variables and the goodness-of-fit tests applied are described. SEM involves four stages: model specification, model estimation, model evaluation and model modification [35]. Under the model specification, measurement and structural equation are considered. The measurement equation is given by

\[ x_i = \Lambda \omega_i + e_i, \quad i = 1, \ldots, n, \]  

---

**Figure 1.** A diagrammatic illustration of the health index using SEM.
where \( x_i \) is an \( p \times 1 \) vector of indicators describing the \( q \times 1 \) random vector of latent variables \( \omega_i \), \( \Lambda \) is \( p \times q \) matrices of the loading coefficients as obtained from the regressions of \( x_i \) on \( \omega_i \) and \( \varepsilon_1 \) is \( p \times 1 \) random vectors of the measurement errors which follow \( N(0, \psi_\varepsilon) \). It is assumed that for \( i = 1, \ldots, n, \omega_i \) is independent, follows a normal distribution \( N(0, \Phi) \) and uncorrelated with the random vector \( \varepsilon_i \).

Let the latent variable \( \omega_i \) be partitioned into \((\eta_i, \xi_i)\), where \( \eta_i \) and \( \xi_i \) are \( m \times 1 \) and \( n \times 1 \) vectors of latent variables, respectively. Equation (2) is the structural equation for explaining the interrelationship among the latent factors in a form of mathematical expression which is given by

\[
\eta_i = B\eta_i + \Gamma\xi_i + \delta_i, \quad i = 1, \ldots, n, \tag{2}
\]

where \( B \) 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, \( \Gamma \) is \( m \times n \) regression parameter matrix for relating the endogenous latent variables and exogenous latent variables, and \( \delta_i \) is \( m \times 1 \) vector of disturbances which is assumed \( N(0, \psi_\delta) \), where \( \psi_\delta \) is a diagonal covariance matrix. It is also assumed that \( \delta_i \) is uncorrelated with \( \xi_i \). Since only one endogenous latent variable is involved in this study, or \( B\eta_i = 0 \), so Equation (2) can be rewritten as \( \eta_i = \Gamma\xi_i + \delta_i \).

The model estimation is to minimize the difference between the hypothesized matrix and the sample covariance matrix based on a suitable fitting function. Since non normal and/or non continuous data are considered in this study, robust-weighted least-square (RWLS) estimation method is an appropriate technique to be utilized. The analysis based on RWLS provides parameter estimates, standard errors, computed \( \chi^2 \) and fit indices which are found using diagonal elements of the weight matrix that are derived from the asymptotic variances of the thresholds and estimates of the latent correlation [10].

The next process in SEM is the model evaluation of the sample parameters. The criteria for estimation of fit include examination of the solution, measure of overall fit and detailed assessment of fit. The first process involved under model evaluation is to check for the appropriateness of each variable such as the right sign or size of the parameter estimates, correlations between parameter estimates, the squared multiple correlations or whether or not the ranges of the standard errors fall within reasonable intervals. The next process is to evaluate the overall model fit by investigating whether or not the specified model fits the data. A complete discussion of model fit is outside the scope of this article, thus, emphasis is given on only three most popular fit indices that are presented by Mplus: the root mean square error of approximation (RMSEA), comparative fit index (CFI) and Tucker Lewis index (TLI) [14,35]. RMSEA provides a measure of the lack of fit of a particular model compared with a perfect or saturated model, where the values of 0.06 or less indicate a good fit, while values larger than 0.10 are an indication of poor fitted models [14]. CFI and TLI compare the improvement of the fit of the proposed model over a more restricted model, called an independence or null model, which specifies no relationships among the variables. The values of CFI and TLI which are closer to 1.0 (or more than 0.80) indicate the better fitted models [14]. Mplus also presents the \( \chi^2 \) and weighted root mean square residual (WRMR). But in this study, neither of these measures is considered as an indicator of the model fit because the \( \chi^2 \) test is sensitive to the sample size and WRMR is an indicator for experimental fit which does not always behave well.

The last step in SEM is the model modification, a process which is done to improve the fit, thereby estimating the most likely relationships between variables. Examples of the methods that could be used in modifying a SEM model are \( \chi^2 \) difference, Lagrange multiplier (ML) and Wald tests. Many programs provide modification indices that indicate the improvement in fit as the result of adding an additional path to the model. For the model modification, Mplus applies the \( \chi^2 \) difference [35].
3.2 Bayesian SEM approach

In this study, the variables gathered are in the form of ordered category. Before conducting the Bayesian analysis, a threshold specification has to be identified in order to treat the ordered categorical data as manifestations of hidden continuous normal distribution. Following is a brief explanation about threshold specification.

Suppose \( X = (x_1, x_2, \ldots, x_n) \) and \( Y = (y_1, y_2, \ldots, y_n) \) be the ordered categorical data matrices and latent continuous variables, respectively. The relationship between \( X \) and \( Y \) is explained by using the threshold specification as follows. For illustration, consider \( x_1 \). The same process can also be done for \( x_2, \ldots, x_n \). Let:

\[
x_1 = c \quad \text{if } \tau_{c-1} < y_1 < \tau_c,
\]

where \( c \) is the number of categories for \( x_1 \), \( \tau_{c-1} \) and \( \tau_c \) denote the threshold levels associated with \( y_1 \). For example, in this study, we consider \( c = 3 \), where \( \tau_0 = -\infty \) and \( \tau_3 = \infty \). Meanwhile, the values of \( \tau_1 \) and \( \tau_2 \) are determined based on the proportion of cases in each category of \( x_1 \) using the formula given by

\[
\tau_k = \Phi^{-1}\left( \sum_{r=1}^{2} \frac{N_r}{N} \right), \quad k = 1, 2,
\]

where \( \Phi^{-1}(\cdot) \) is the inverse of the standardized normal distribution, \( N_r \) is the number of cases in the \( r \)th category and \( N \) is the total number of cases. Here, it is assumed that \( y_1 \) follows a normal distribution. Thus, we have \( Y = (y_1, y_2, \ldots, y_n) \) following a multivariate normal distribution.

Under the Bayesian SEM, we also consider \( X = (x_1, x_2, \ldots, x_n) \) and \( Y = (y_1, y_2, \ldots, y_n) \) to be the ordered categorical data matrices and latent continuous variables, respectively, and \( \Omega = (\omega_1, \omega_2, \ldots, \omega_n) \) be the matrix of latent variables. The observed data \( X \) are augmented with the latent data \( (Y, \Omega) \) in the posterior analysis. In this subsection, we will apply the Bayesian estimation in SEM in order to do posterior analysis to obtain the values of unknown threshold in \( \tau = (\tau_1, \tau_2) \), joint Bayesian estimates of \( \Omega \) and the structural parameter \( \theta \), a vector that includes all the unknown parameters in \( \Phi, \psi, \psi_e, \Lambda, \Lambda_\omega \).

The Bayesian method is applied to derive the posterior distribution \( \{ \tau, \theta, \Omega \vert X \} \). The user may apply maximum likelihood (ML) estimation method to derive the posterior distribution. However, a high-dimensional integration is required; thus, it is difficult to apply an ML estimation method in this study [3,19,21]. Thus, the Gibbs sampler algorithm is applied to overcome this difficulty. The Gibbs sampler is a Markov Chain Monte Carlo (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 [33]. Since ordered categorical variables are used, the latent matrix \( Y \) should be augmented in the posterior analysis, resulting in the joint posterior distribution \( \{ \tau, \theta, \Omega \vert Y \} \). The Gibbs sampler process starts with the setting of initial starting values \( (\tau^{(0)}, \theta^{(0)}, \Omega^{(0)}, Y^{(0)}) \), and then conduct the simulation for \( (\tau^{(1)}, \theta^{(1)}, \Omega^{(1)}, Y^{(1)}) \). At the \( r \)th iteration, by making use of the current values \( (\tau^{(r)}, \theta^{(r)}, \Omega^{(r)}, Y^{(r)}) \), the Gibbs sampler is carried out as follows:

a: Generate \( \Omega^{(r+1)} \) from \( p(\Omega \vert \tau^{(r)}, \theta^{(r)}, Y^{(r)}, X) \),
b: Generate \( \theta^{(r+1)} \) from \( p(\theta \vert \Omega^{(r+1)}, \tau^{(r)}, Y^{(r)}, X) \),
c: Generate \( (\tau^{(r+1)}, Y^{(r+1)}) \) from \( p(\tau, Y \vert \Omega^{(r+1)}, \theta^{(r+1)}, X) \).

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 is discussed in Lee and Shi [19] or Lee [18]. In the process when determining the posterior distribution, the
As given in Figure 1, the hypothesized model in this study consists of 14 indicator variables with information for the fit of the model. The residuals estimates for measurement equation and structural equation, we plot the residual estimates versus latent variable estimates to give deviation [6].

The next process in the Bayesian SEM is a 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 and within multiple chains, denoted by $R$. The estimated parameters converge if the value of $R$ 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 ($\hat{\epsilon}_i$) can be obtained from

$$\hat{\epsilon}_i = y_i - \hat{\Lambda}\hat{\xi}_i, \quad i = 1, \ldots, n,$$

where $\hat{\Lambda}$ and $\hat{\xi}_i$ 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 estimated parameters converge if the value of $R$ 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 ($\hat{\epsilon}_i$) can be obtained from

$$\hat{\epsilon}_i = y_i - \hat{\Lambda}\hat{\xi}_i, \quad i = 1, \ldots, n,$$

where $\hat{\Lambda}$ and $\hat{\xi}_i$ 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 ($\hat{\delta}_i$) can be obtained from following equation:

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

where $\hat{B}$, $\hat{\eta}_i$, $\hat{\xi}_i$ and $\hat{\Gamma}$ 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.

### 3.3 Modeling description

As given in Figure 1, the hypothesized model in this study consists of 14 indicator variables with three exogenous latent variables and one endogenous latent variable. The measurement equation is defined by

$$y_i = \Lambda \omega_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\omega_i = (\eta_1, \xi_{i1}, \xi_{i2}, \xi_{i3})^T$. It is assumed that $\epsilon_i$ is independent, following $N(0, \psi_\epsilon)$ and uncorrelated with the latent variable $\omega_i$. The structural equation is modeled as follows:

$$\eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i3} + \delta_i,$$

where $(\xi_{i1}, \xi_{i2}, \xi_{i3})^T$ is distributed as $N(0, \Phi)$ and independent with $\delta_i$ which is distributed as $N(0, \psi_\delta)$. 
In the data analysis, we use Mplus version 5.2 [23] to estimate the parameters for the classical model, while the Bayesian model is fitted to the data using winBUGS version 1.4 [33]. The hierarchical structure is implemented by selecting the prior information for parameters involved in the hypothesized model as described in Equations (5)–(7).

4. Results

The results of model fitting obtained based on the classical SEM approach are provided in Figure 2. These indicate that the proposed model fits the data reasonably well since RMSEA = 0.053, CFI = 0.802 and TLI = 0.772. The values of RMSEA and CFI are well within the acceptable range and the value of TLI is on the borderline since the range of acceptable value of TLI is between 0.8 and 1.0 [14]. In addition, we have almost 100% power of rejecting the hypothesis that the model does not fit our data at the 0.05 significance level, decided based on the value of RMSEA found [22]. The calculation of the value of RMSEA is based on the computed $\chi^2$ value of the model. Since the sample size in this current study is large enough, the power of the test will ensure that the hypothesis is rejected.

In the Bayesian analysis, it is important to test the sensitivity of the Bayesian analysis with respect to the choice of the priors. In order to achieve this goal, we consider the model comparison with three types of prior inputs. In assigning the values for hyper parameter, we take small variance for each parameter, since we have confidence to have good prior information about a parameter [18]. We also follow the suggestion of Kass and Raftery [16] to perturb the hyperparameters as follows. For all three prior inputs, fix $\alpha_{0k} = \alpha_{0\tilde{k}} = 10$, and $\beta_{0k} = \beta_{0\tilde{k}} = 8$ and those corresponding to $\Phi$ are $\rho_0 = 30, R_0^{-1} = 8I, H_{0\tilde{k}} = 0.1I$, and $H_{0\tilde{k}k} = 0.1I$. Accordingly, the three prior inputs are summarized as follows:

(a) Type I prior: the unknown loadings in $\Lambda$ are all taken to be 0.5, those values corresponding to $\{\gamma_1, \gamma_2, \gamma_3\}$ are $\{0.7, 0.5, 0.1\}$.
(b) Type II prior: the hyperparameters are equal to half of the values given in (a).
(c) Type III prior: the hyperparameters are equal to twice of the values given in (a).

The results found based on the three type of prior inputs are presented in Table 1.

---

Figure 2. The fitted model of the Classical SEM, which includes the standardized parameter estimates, covariance among latent variables and their standard errors in parentheses and measurement error of the model.
We can see from Table 1 that the parameter estimates and standard errors obtained under various prior inputs are reasonably close. We could conclude here that the statistics found based on the Bayesian SEM is not sensitive to these three different prior inputs or we can also say that the Bayesian SEM applied here is quite robust to the different prior inputs. Accordingly, for the purpose of discussion of the results found using the Bayesian SEM, we will use the results obtained using Type I prior. These results are provided in Figure 3.

Test of convergence statistics for all parameters of interest are plotted and we found that the values of $R$ are close to 1. Plots of sequences of observations corresponding to some parameters generated by two different initial values are also examined. All plots indicate that the algorithm converged in less than 7000 iteration. In Figure 4, we present several plots for illustrative purposes. The Monte Carlo errors obtained for all the parameters considered are also less than 5% of the sample standard deviation.

Plots of the estimated residual versus the case number are also checked to assess the plausibility of the proposed model. All plots lie within two parallel horizontal lines, centered at zero and no
trends are detected. We could conclude here that the estimated model which is obtained based on the Bayesian SEM analysis would be considered adequate and could be acceptable.

The next analysis is the simulation study using the Bootstrap technique. The goal of simulation study here is to reveal the performance of the Bayesian approach and its associated algorithm in recovering the true parameters. Simulation study does so by generating a set of new data set by sampling with replacement from the original data set, and fitting the model to each new data set. To compute standard errors for calculating the 95% confidence interval of all parameters in this study, roughly 100 model fits are determined. Table 2 presents the result taken from the simulation study.

Table 2 shows that all parameter estimates fall within the 95% bootstrap percentile intervals obtained from the simulation study. Bootstrap percentile intervals seem to work well here. It means that the estimated posterior mean are acceptable. Thus, we believe that the power of our Bayesian SEM could yield the best fit for the model.

Based on Figures 2 and 3, we obtain the estimated structural equations that address the relationship between the health index with socio-demography, mental health and lifestyle for the classical SEM and the Bayesian SEM which are given by

\[
\hat{\eta}_{\text{clasSEM}} = 1.248 \xi_1 + 0.724 \xi_2 - 0.033 \xi_3 \tag{12}
\]

and

\[
\hat{\eta}_{\text{BaySEM}} = 0.894 \xi_1 + 0.484 \xi_2 - 0.039 \xi_3, \tag{13}
\]

respectively.

These estimated structural equations indicated that socio-demographic status (\(\xi_1\)) has the greatest effect on the health index (\(\eta\)) than the other two latent variables. The relationship between socio-demography and health index is significant. One can conclude here that socio-demographic status is significantly correlated to the health condition, which implies that people of good

Figure 3. The fitted model of the Bayesian SEM, which includes the standardized parameter estimates, covariance among latent variables and their standard errors in parentheses and measurement errors of the model.
socio-demographic status tend to experience a better health condition. This study also finds that lifestyle has a direct effect on the health index and this relationship is statistically significant. It is also obtained that mental health has no significant relationship to the health index. It is possible that there are some other important variables which should also be incorporated into the model as
Table 2. Simulation results using the bootstrap technique and posterior mean.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bootstrap mean</th>
<th>Bootstrap standard deviation</th>
<th>95% bootstrap percentile interval</th>
<th>Estimated posterior mean using HRA data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>0.878</td>
<td>0.064</td>
<td>0.751 – 1.005</td>
<td>0.894</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.456</td>
<td>0.067</td>
<td>0.323 – 0.589</td>
<td>0.484</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>−0.038</td>
<td>0.051</td>
<td>−0.139 – 0.062</td>
<td>−0.039</td>
</tr>
<tr>
<td>$\lambda_{21}$</td>
<td>0.848</td>
<td>0.042</td>
<td>0.765 – 0.932</td>
<td>0.842</td>
</tr>
<tr>
<td>$\lambda_{31}$</td>
<td>−0.800</td>
<td>0.051</td>
<td>−0.902 – −0.699</td>
<td>−0.794</td>
</tr>
<tr>
<td>$\lambda_{52}$</td>
<td>0.522</td>
<td>0.064</td>
<td>0.396 – 0.648</td>
<td>0.529</td>
</tr>
<tr>
<td>$\lambda_{62}$</td>
<td>1.343</td>
<td>0.068</td>
<td>1.209 – 1.477</td>
<td>1.338</td>
</tr>
<tr>
<td>$\lambda_{72}$</td>
<td>0.247</td>
<td>0.048</td>
<td>0.152 – 0.341</td>
<td>0.218</td>
</tr>
<tr>
<td>$\lambda_{93}$</td>
<td>0.573</td>
<td>0.062</td>
<td>0.451 – 0.695</td>
<td>0.548</td>
</tr>
<tr>
<td>$\lambda_{103}$</td>
<td>1.290</td>
<td>0.076</td>
<td>1.139 – 1.440</td>
<td>1.271</td>
</tr>
<tr>
<td>$\lambda_{124}$</td>
<td>0.436</td>
<td>0.039</td>
<td>0.359 – 0.513</td>
<td>0.432</td>
</tr>
<tr>
<td>$\lambda_{134}$</td>
<td>0.370</td>
<td>0.045</td>
<td>0.281 – 0.459</td>
<td>0.363</td>
</tr>
<tr>
<td>$\lambda_{144}$</td>
<td>0.178</td>
<td>0.032</td>
<td>0.114 – 0.242</td>
<td>0.165</td>
</tr>
<tr>
<td>$\Phi_{11}$</td>
<td>0.489</td>
<td>0.035</td>
<td>0.419 – 0.559</td>
<td>0.488</td>
</tr>
<tr>
<td>$\Phi_{12}$</td>
<td>−0.230</td>
<td>0.023</td>
<td>−0.276 – −0.185</td>
<td>−0.234</td>
</tr>
<tr>
<td>$\Phi_{13}$</td>
<td>−0.099</td>
<td>0.011</td>
<td>−0.121 – −0.076</td>
<td>−0.094</td>
</tr>
<tr>
<td>$\Phi_{22}$</td>
<td>0.283</td>
<td>0.024</td>
<td>0.234 – 0.331</td>
<td>0.284</td>
</tr>
<tr>
<td>$\Phi_{23}$</td>
<td>0.089</td>
<td>0.008</td>
<td>0.072 – 0.107</td>
<td>0.091</td>
</tr>
<tr>
<td>$\Phi_{33}$</td>
<td>0.244</td>
<td>0.025</td>
<td>0.194 – 0.293</td>
<td>0.243</td>
</tr>
<tr>
<td>$\Psi_3$</td>
<td>0.348</td>
<td>0.028</td>
<td>0.292 – 0.403</td>
<td>0.348</td>
</tr>
<tr>
<td>$\Psi_{31}$</td>
<td>0.865</td>
<td>0.104</td>
<td>0.660 – 1.069</td>
<td>0.870</td>
</tr>
<tr>
<td>$\Psi_{32}$</td>
<td>0.675</td>
<td>0.028</td>
<td>0.620 – 0.731</td>
<td>0.673</td>
</tr>
<tr>
<td>$\Psi_{33}$</td>
<td>0.692</td>
<td>0.029</td>
<td>0.633 – 0.750</td>
<td>0.690</td>
</tr>
<tr>
<td>$\Psi_{34}$</td>
<td>0.745</td>
<td>0.036</td>
<td>0.674 – 0.816</td>
<td>0.760</td>
</tr>
<tr>
<td>$\Psi_{52}$</td>
<td>0.931</td>
<td>0.043</td>
<td>0.846 – 0.915</td>
<td>0.920</td>
</tr>
<tr>
<td>$\Psi_{62}$</td>
<td>0.504</td>
<td>0.032</td>
<td>0.442 – 0.567</td>
<td>0.505</td>
</tr>
<tr>
<td>$\Psi_{72}$</td>
<td>0.982</td>
<td>0.052</td>
<td>0.879 – 1.085</td>
<td>0.987</td>
</tr>
<tr>
<td>$\Psi_{83}$</td>
<td>0.780</td>
<td>0.032</td>
<td>0.716 – 0.844</td>
<td>0.774</td>
</tr>
<tr>
<td>$\Psi_{93}$</td>
<td>0.916</td>
<td>0.039</td>
<td>0.840 – 0.993</td>
<td>0.926</td>
</tr>
<tr>
<td>$\Psi_{103}$</td>
<td>0.594</td>
<td>0.035</td>
<td>0.525 – 0.664</td>
<td>0.606</td>
</tr>
<tr>
<td>$\Psi_{114}$</td>
<td>0.431</td>
<td>0.033</td>
<td>0.366 – 0.497</td>
<td>0.425</td>
</tr>
<tr>
<td>$\Psi_{124}$</td>
<td>0.888</td>
<td>0.036</td>
<td>0.817 – 0.959</td>
<td>0.886</td>
</tr>
<tr>
<td>$\Psi_{134}$</td>
<td>0.912</td>
<td>0.038</td>
<td>0.837 – 0.987</td>
<td>0.922</td>
</tr>
<tr>
<td>$\Psi_{144}$</td>
<td>0.972</td>
<td>0.033</td>
<td>0.907 – 1.037</td>
<td>0.983</td>
</tr>
</tbody>
</table>

Note: HRA, Health Risk Assessment.

indicators for mental health, such as suffering from schizophrenia, depression, anxiety or social disorder [11,36].

The values of the unstandardized coefficient of factor loading and the associated standard errors for each indicator variable in the measurement equations obtained based on both approaches are presented in Table 3.

It is clear from Table 3 that both models yield almost identical estimates of the factor loading. All indicator variables that we hypothesized as predictors are significantly related to their respective latent factors. It is interesting to observe that standard errors for the parameter estimates found under the Bayesian SEM are generally slightly smaller than those found based on the classical SEM. Table 3 also shows that the length of the 95% confidence intervals associated with the parameters obtained from the Bayesian SEM are generally shorter compared with those of the classical SEM. This is not surprising due to the extra information brought by the prior distribution.
Table 3. Coefficient regressions in measurement model.

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Indicator variable</th>
<th>Estimate</th>
<th>Classical SEM (SE) 95% CI</th>
<th>Bayesian SEM (SE) 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio demography (ξ₁)</td>
<td>Employment (X₁₁)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Education (X₁₂)</td>
<td>0.901 (0.041)*</td>
<td>0.845 (0.079)*</td>
<td>(0.821, 0.980)</td>
</tr>
<tr>
<td></td>
<td>Age group (X₁₃)</td>
<td>−0.946 (0.041)*</td>
<td>−0.796 (0.084)*</td>
<td>(−1.028, −0.865)</td>
</tr>
<tr>
<td>Lifestyle (ξ₂)</td>
<td>Smoking (X₄₂)</td>
<td>0.675 (0.065)*</td>
<td>0.529 (0.058)*</td>
<td>(0.548, 0.802)</td>
</tr>
<tr>
<td></td>
<td>Exercise (X₅₂)</td>
<td>1.556 (0.114)*</td>
<td>1.342 (0.088)*</td>
<td>(1.334, 1.779)</td>
</tr>
<tr>
<td></td>
<td>Working hours (X₆₂)</td>
<td>0.131 (0.059)*</td>
<td>0.220 (0.059)*</td>
<td>(0.016, 0.246)</td>
</tr>
<tr>
<td>Mental health (ξ₃)</td>
<td>Happy (X₈₃)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Problem (X₉₃)</td>
<td>0.661 (0.099)*</td>
<td>0.543 (0.084)*</td>
<td>(0.467, 0.856)</td>
</tr>
<tr>
<td></td>
<td>Stress level (X₁₀₃)</td>
<td>2.583 (0.583)*</td>
<td>1.278 (0.114)*</td>
<td>(1.439, 3.726)</td>
</tr>
<tr>
<td>Health index (η)</td>
<td>BP (X₁₁₄)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>BMI (X₁₂₄)</td>
<td>0.429 (0.039)*</td>
<td>0.432 (0.034)*</td>
<td>(0.352, 0.505)</td>
</tr>
<tr>
<td></td>
<td>Cholesterol (X₁₃₄)</td>
<td>0.478 (0.041)*</td>
<td>0.362 (0.037)*</td>
<td>(0.398, 0.558)</td>
</tr>
<tr>
<td></td>
<td>Health problem (X₁₄₄)</td>
<td>0.215 (0.039)*</td>
<td>0.165 (0.033)*</td>
<td>(0.138, 0.293)</td>
</tr>
</tbody>
</table>

Notes: BP, blood pressure; BMI, body mass index; SE, standard error; CI, confidence interval.
*Significant at the 5% level.

In our Bayesian analysis, we have used the conjugate prior distribution for updating the current information on the parameter. In the case of no prior information, it has been argued that it is better to use non-informative prior inputs rather than bad subjective prior inputs [20]. In this study, however, we have a large sample, thus, the estimated parameters obtained are found to be less sensitive to the different choices of the prior inputs considered. Thus, prior inputs should be selected with great care, particularly when the sample size is small.

5. Discussion

The main purpose of the present study is to demonstrate the value of the classical SEM and the Bayesian SEM technique in modeling the health index of an individual in Hulu Langat. Under the classical perspective, CFA SEM is used to test the appropriate number of latent variables for explaining the observed items. The strength of SEM is its ability to do a simultaneous test in order to describe the relationship between the observed variables and the respective latent variables as well as the relationship among the latent variables [35]. The analysis in this study is implemented under Mplus version 5.2, a flexible tool which allows one to examine the relationship involving the violation of normal assumption of the variables considered in the model. In addition, for comparison with the results under the classical approach, the Bayesian SEM is applied using winBUGS version 1.4.

Even though many works have been done on determining the health index, not much works have done on modeling of this index using SEM, particularly when information on socio-demography, lifestyle, mental health and biomarkers are considered. The indicators that are found significant
in explaining the latent factors considered in this study are as follows. The socio-demographic indicators are employment status, education level and age-group. Lifestyle is explained using smoking habit, frequency of engaging in physical exercise, number of working hours and number of sleeping hours. Stress levels, how do the respondent feel about her/his life and experience of serious problems that respondents have, are used as indicators to measure mental health condition. This study found that socio-demography and lifestyle have a significant effect on the health index, but mental health does not. These findings are similar to the study of Rizal [30], who indicated that hypertension, which he considered as an indicator of a health index, is significantly related to indicators of socio-demography, such as age, education level, employment status and indicators of lifestyle, i.e. smoking habit and exercise. He also finds that mental health does not have a significant effect on the health index. This result is similar to the study of Jamsiah et al. [15], which found that hypertension have no association with stress, which they believe to be an indicator of mental health.

However, some studies found that mental health is significantly related to the health index [24]. This conflicting evidence between this study and some other previous studies could possibly be due to the choice of indicators used to explain mental health. Some studies have considered schizophrenia, depression and anxiety for describing mental health [11,36]. However, for the majority of Malaysians, we believe, and it remains to be investigated, that it is uncalled-for in our society for someone to admit of having a mental disease such as schizophrenia because having schizophrenia is associated with being insane and people that are insane have no place in the society. Accordingly, distant proxies such as the level of stress experienced by the respondents have been used in this study instead of medical diagnoses as indicators of mental health. In addition, the studies by Nakayama et al. [25] and Nagano et al. [24], which were carried out in the factories indicate that the Japanese workers who used to work overtime have a high level of stress, implying that long working hours contribute a negative impact on mental health. Thus, length of working hours is another factor which could possibly be considered as an indicator for mental health.

It is envisaged that factors such as living environment, food consumption and length of working hours should also be considered when modeling the health index. The study conducted by Kiyu et al. [17] in Sarawak, for example, has identified that smoking habit, exercise habit and living environment are factors that are significant in determining the level of health. Noor [27] has indicated in his work that food consumption and lifestyle play an important role in influencing the level of health of an individual. Certainly, many other factors in addition to these three factors could possibly influence a person’s health condition.

The idea of modeling the health index by considering various indicators which describe the latent factors could further be explored by incorporating new HRA survey data. This idea is particularly suitable under the sequential Bayesian approach by considering results of this study as the prior input for the new survey. In this way, at least to some extent, the current health status of individual living in Hulu Langat can be monitored.

Acknowledgements

The authors thank the Associate Professor Khalib A Latiff and Mr Khairul Yusof from the Department of Community Health, Medical Faculty, UKM, who furnished us the health survey data used in this study. This research was financially supported by Directorate General of Higher Education, Ministry of National Education, Indonesia. In addition, a partial support was also obtained from UKM under the research grant with the code UKM-ST-06-FRGS0011-2007. We also thank several anonymous referees for their constructive comments which have improved the final version of this paper.

References


