Quantile Regression Approach to Determine The Indicator of Health Status

Ferra Yanuar

Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Andalas, Kampus Limau Manis, 25163, Padang – Indonesia,
Email: ferrayanuar@yahoo.co.id

***Abstract*- The aim of this study are to demonstrate the application of the quantile regression technique to determine the indicators of the health status and to identify the characteristics of healthy and unhealthy individual. Quantile regression approach was applied to 530 health survey data of respondents from Hulu Langat, Malaysia. The data survey informs about some indicator variables that relate to socio-demographic status, lifestyle, mental health and value of health status of the respondents. It is assumed that employment level, education level, age-group, smoking habits, physical exercise, average working and sleeping in a day, stress level, happiness in life and social problems that she/he has have effect to the health status of related individual. All ten indicator variables are in ordered categorical type of data and health status value as the response is in continuous type. This study found that at the middle quantile, employment level, education level, age-group and smoking habits are significant indicators of health status. This study also identify the characteristics of healthy individuals who are at the upper quartile level of health are individuals working, educated individuals, young and non-smoking. Meanwhile, the characteristics of unhealthy individuals who are at the lower quartile level of health are highly educated and older than 30 years old.**

***Index Terms*— quantile regression, health status, healthy individual, unhealthy individual, ordered categorical.**

I. INTRODUCTION

Health status of an individual is the health condition that had by corresponding individual. The indicators of individual's health status could be different to other individual and those could change from time to time. It is important to identify the determinants of health status since the information on it could be used by many parties. Information of health status of an individual that is often gathered in a health survey of which can be used by decision makers to allocate resources prudently when plans on activities which aim at improving the overall health status of the community where she or he lives.

Various studies have indicated that many factors could influence the health status of an individual. Ramirez (2004), Broadman (2004), Shi (1998) and Uitenbroek et al. (1996) proved that education level, employment level and age group have direct affect to the health status. Broadman (2004), Nakayama et al. (2001), Uitenbroek et al. (1996), Cheadle et al. (1994) have shown that smoking habits, physical exercise, average sleeping hours and average working hours in a day also affect the health status of an individual. Meanwhile level of stress, happiness in life and the number of serious problems that was faced during the last year are proved by Nakayama et al. (2001) and Boardman (2004) as factors of health status as well. Yanuar, et al. (2010) have been used all indicator variables mentioned above to construct the health status index using structural equation modeling approach.

Based on those literatures, the indicator variables which assumed have relation to the health status in this present study are education level, employment level, age group, smoking habits, physical exercise, average sleeping hours and average working hours in a day, level of stress, happiness in life and the number of serious problems that was faced during the last year.

The main purpose of this study is to determine the significant indicator variables that affect the health status of an individual. We construct the model by applying the data of National Health Morbidity Survey (NHMS) III in Malaysia. Then, this study also investigate the characteristics of healthy and unhealthy individuals. Thus, the quantile regression approach is applied here as an appropriate technique in achieving the purposes of the study. Quantile regression is a statistical method used to estimate models based on conditional quantile function. Quantile regression models the relation between a set of predictor variables and specific quantiles of the response variable. In this quantile regression modeling, the measures of health status are regressed on the all ten factors as mentioned above. If the value of health status is at lower quantile (under 30%), it indicates that corresponding individuals are in unhealthy condition, contrary if the value of health status is at higher quantile (more than 70%) it means that the individuals are in healthy condition.

The advantages of quantile regression, relative to the ordinary least squares regression, are that the quantile regression estimates are more robust against outliers in the response measurements (Koenker, 2000) and more informative (Koenker & Hallock, 2001). In practice we often prefer using different measures of central tendency and statistical dispersion to obtain a more comprehensive analysis of the relationship between variables.

II. GENERATION OF THE DATA

This present study apply the analysis on a part of data set obtained from 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 [10]. 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. The independent variables choosen to be included in the analysis are shown in the following Table 1.

Table 1. The distribution of independent variables and its categories

No	Variables	Categories	Code	Frequencies	Percent
1	Employment level	Unemployed	1	147	27.7
		Ordinary	2	242	45.7
		Professional	3	141	26.6
2	Education level	Low	1	109	20.6
		Middle	2	319	60.2
		High	3	102	19.2
3	Age-group	Between 18 to 30 years old	1	180	34.0
		Between 31 to 56 years old	2	330	62.3
		More than 56 years old	3	20	3.8
4	Smoking habit	Smoker	1	84	15.8
		Ever smoker	2	67	12.6
		Non smoker	3	379	71.5
5	Physical exercise	None	1	283	53.4
		1 or 2 times in a week	2	151	28.5
		More than 2 times in a week	3	96	18.1
6	Working hours	More than 14 hours	1	25	4.7
		9-14 hours	2	186	35.1
		Less than 9 hours	3	319	60.2
7	Sleeping hours	Less than 7 hours	1	69	13.0
		More than 8 hours	2	124	23.4
		7-8 hours	3	337	63.6
8	Social problem	High	1	4	0.8
		Middle	2	319	60.2
		Low	3	207	39.1
9	Happiness in Life	Not happy	1	38	7.2
		Average	2	155	29.2
		Happy	3	337	63.6
10	Stress level	High	1	19	3.6
		Middle	2	199	37.5
		Normal	3	312	58.9

Meanwhile the dependent variable in this quantile regression model is health status. The measure of health status is derived by using Bayesian structural equation modeling. The value of health status is treated as the normal variate and assumed to lie between -3 to 3, where steps and process of measuring as were done by Yanuar et al. (2013). Therefore, the value of health status is in continuous type and described in this following Table 2 and Graph 1.

Table 2. Description The Distribution of Health Status

N		530
Mean		0.011
Median		0.080
Mode		-0.224
Minimum		-1.684
Maximum		0.966
τth quantile =	10	-0.787
	20	-0.477
	25	-0.338
	30	-0.268
	40	-0.153
	50	0.080
	60	0.164
	70	0.337
	75	0.539
	80	0.654
	90	0.744

Table 2 above informs us that the range of health status's value is between -1.684 to 0.966. It means that the health status's value of the healthies individual is 0.966, meanwhile the unhealthies individual is -1.684. The median of health status is 0.080 which presented at the 0.5th quantile. We can interprete here that half of respondents who are involved in the NHMS III in Hulu Langat, the value of their health status are less than 0.080 and others are more than 0.080. The Table 1 also shows that 30% of respondents have their health status's value less than -0.268 and 70% of them are more than -0.268.

The value of health status here is categorized into three groups. First is the individuals who their health status'value are distributed at less than 0.3th quantile and categorized them as 'unhealthy' group. Then, the health status's value are between 0.3th to 0.7th quantile categorized as 'middle' group and more than 0.7th quantile are classified as 'healthy' group.

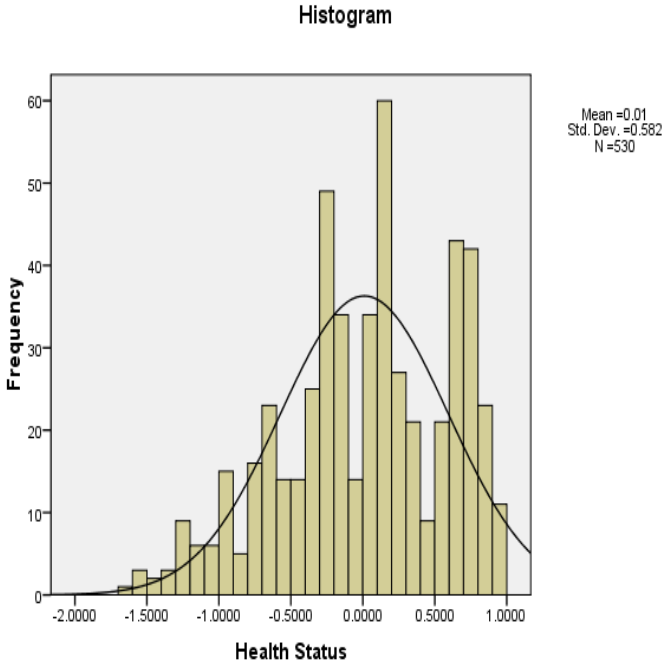

Figure 1. Bar graph of health status value.

III. THEORETICAL MODEL OF QUANTILE REGRESSION FOR HEALTH STATUS

In this paper we use the quantile regression approach to model the health status on the ten independent variables. The distribution of health status for τth quantile have been shown in Table 2 above. The τth quantile of ordered data may be defined as the value of the τ fraction of the data below it and 1-τ fraction of the data above it where 0<τ<1. Here, we will identify the significant independet variables which have relation to the health status for each corresponding τth quantile. Median that equal to 0.5th quantiles is a commonly used quantile. The estimation method used in quantile regression is based on minimizing sums of squared residuals or written as:

Assumed *i*'s health status of individual denoted as H_i, the quantile regression model pertaining to the τth quantile can be formulated as the following:

$$H_i = \beta_\tau{}' X_i + e_i, \quad i = 1,2,...,n \qquad (1)$$

Where β_τ is $1 \times p$ vector of coefficient, X_i is $p \times 1$ vector of independent variables, e_i is *i*th residual. Meanwhile the conditional quantile of H_i given $X_i$or $Q_\tau(H_i|X_i)$ could be formulated as below :

$$Q_\tau(H_i|X_i) = \beta_\tau{}' X_i, \qquad (2)$$

The coefficients are estimated by minimizing:

$$\widehat{\beta}(\tau) = \min_{\beta_\tau} \sum_i Q_\tau(H_i - \beta_\tau{}' X_i|X_i) \qquad (3)$$

For any τ in the interval (0,1), the equation (3) is equivalent to the following equation:

$$\hat{\beta}(\tau) = \min_{\beta_\tau} \{\sum_{i=1}^{n} \tau|H_i - \beta_\tau{}' X_i| + \sum_{i=1}^{n}(1-\tau)|H_i - \beta_\tau{}' X_i|\}, \qquad 0 < \tau < 1 \qquad (4)$$

Therefore, equation (4) can also be interpreted with the estimation of τth sample quantile. For a given p regressors X_i, here we estimate nine quntiles $\{\beta_\tau, \tau = 0.10, 0.20, \ldots, 0.90\}$ of coefficients. In this present study, the vector of independents variables, X_i, contains the indicators which assumed related to the health status. The distribution of independent variables and its categories have been presented in Table 1. We have seen that all indicators take ordinal integer value, with ranging from 1 to 3. For each indicator, code 1 is the lowest category and posited as reference indicator for other two categories.

To implement this quantile regression technique, the researchers can use various software. The most popular software to be applied are R-Program (Verzani, 2005) and S-Plus (Yu, 2003). This present study uses R version 2.14.1 to model the health status of an individual.

IV. RESULTS

In this quantile regression, we regress simultaneously the health status on some potential indicator variables, such as employment level, education level, age-group, smoking habits, physical exercise in a week, average working per day, average sleeping per day, social problems, happiness in life and stress level.

After applying the multivariate quantile regression on the hypothesis model using NHMS III data, we found that the indicator variables which significantly have relationship on the health status distribution are employment level, education level, age-group and smoking habits. Table 3 below presents the result of quantile regression analysis at the different quantile, for τ = (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90).

Table 3. The Value of Coefficient Regressions for Each Independent Varibales (Standard Error)

τth Quantile	Intercept	Independent Variables							
		Ordinary worker	Professional worker	Middle educated	Highly educated	31-56 years old	More than 56 years old	Ever smoker	Non smoker
0.1	-0.483* (0.229)	0.166* (0.085)	0.069 (0.163)	-0.046 (0.084)	0.267** (0.110)	-0.532 ** (0.079)	-0.637** (0.259)	0.105 (0.215)	-0.007 (0.212)
0.2	-0.331** (0.105)	0.070 (0.087)	0.105 (0.077)	0.059 (0.079)	0.383** (0.087)	-0.432** (0.061)	-0.444 (0.268)	-0.023 (0.108)	-0.017 (0.089)
0.3	-0.204 (0.114)	0.287** (0.078)	0.116 (0.087)	0.043 (0.069)	0.371** (0.084)	-0.440** (0.063)	-0.613** (0.072)	0.020 (0.092)	0.018 (0.058)
0.4	0.030 (0.130)	0.072 (0.074)	0.123 (0.093)	0.058 (0.079)	0.156* (0.090)	-0.432** (0.057)	-0.715** (0.133)	0.003 (0.110)	0.013 (0.093)
0.5	0.165 (0.129)	0.198** (0.083)	0.101 (0.088)	0.066 (0.080)	0.296** (0.121)	-0.330** (0.072)	-0.541** (0.194)	-0.107 (0.123)	-0.107 (0.091)
0.6	0.263** (0.109)	0.223** (0.072)	0.114 (0.089)	0.073 (0.051)	0.322** (0.075)	-0.433** (0.080)	-0.663** (0.121)	0.003 (0.089)	0.011 (0.047)
0.7	0.553** (0.120)	0.072 (0.083)	0.104 (0.110)	0.078 (0.082)	0.223** (0.084)	-0.553** (0.062)	-0.762** (0.127)	0.014 (0.093)	0.002 (0.066)
0.8	0.550** (0.025)	0.072** (0.020)	0.153** (0.032)	0.091** (0.021)	0.221** (0.032)	-0.102 (0.081)	-0.659** (0.096)	0.014 (0.020)	0.016 (0.018)
0.9	0.560** (0.003)	0.072** (0.004)	0.160** (0.006)	0.092** (0.003)	0.205** (0.007)	-0.063** (0.003)	-0.535** (0.093)	0.004 (0.004)	0.024** (0.004)

** p-value ≤ 0.01, * p-value ≤ 0.05

i Based on table 3 above, we could determine that the health status model for individual living at Hulu Langat using multivariat quantile regression approach is modeled as following:

Q_τ (Health status) = $\beta_0 + \beta_1$ Ordinary worker + β_2 Professional worker +β_3 Middle educated + β_4 Highly educated + β_5 31-56 years old + β_6 More than 56 years old + β_7 Ever smoke + β_8 Non smoking (5)

where:

Q_τ = Quantile regression equation at the τth quantiles = (0.10, 0.20, …, 0.90) of n data
β_0 = Intercept at the τth quantile
β_1 = Regression coeficient of ordinary worker at the τth quantile
β_2 = Regression coeficient of professional worker at the τth quantile
β_3 = Regression coeficient of middle educated at the τth quantile
β_4 = Regression coeficient of highly educated at the τth quantile
β_5 = Regression coeficient of 31-56 years old at the τth quantile
β_6 = Regression coeficient of more than 56 years old at the τth quantile
β_7 = Regression coeficient of ever smoke at the τth quantile
β_8 = Regression coeficient of non smoking at the τth quantile

Table 3 also present the coefficient estimation for each significant indicator variables as the results of multivariate quantile regression analysis. With respect to employment status indicate positive effects on almost all quantiles of the health status distribution. The individual who are in ordinary level of her/his employment significantly have higher health status than unemployed on all quantiles except at 0.2, 0.4 and 0.7th quantile. Meanwhile the professional worker tends to have higher health status level significantly than unemployed at the upper quantile, i.e at 0.8 and 0.9th quantile.

Regarding to personal education level, consistent with findings in the relevant literature, positively effects the health status on all quantiles. The individual with highly educated would have better health condition than low education level on all quantiles. The

respondents with middle education level proven also significantly have higher health status value than low education level at the upper quantile, such as at 0.8 and 0.9th quantile.

Relating to the age-group indicator variable, this present study found that age group at more than 31 years old gave negative affect to the health status, marked by negative sign for all coefficient regression on all quantiles for age group between 31 to 56 year old and more than 56 years old. The respondents at 31 to 56 years old tends to have lower health status value compare to respondents at less than 31 years old. The respondents at more than 56 years old would have worst health condition than respondents at less than 31 years old on all quantiles. We suspect here that as age increases the health status condition would decreases. It means that older individual would have lower health status value compare to the younger individual.

With respect to the smoking habits, we found the non smoker respondents tend to have better health status condition significantly compare to smoker respondents at 90th quantile only.

The pattern of health status model using multivariate quantile regression could be drawn in a graph as presented in Figure 2 below. Horizontal dashed lines represents the classical ordinary least square regression line using mean square estimation, meanwhile horizontal line appear in some graphs indicate line at zero point.

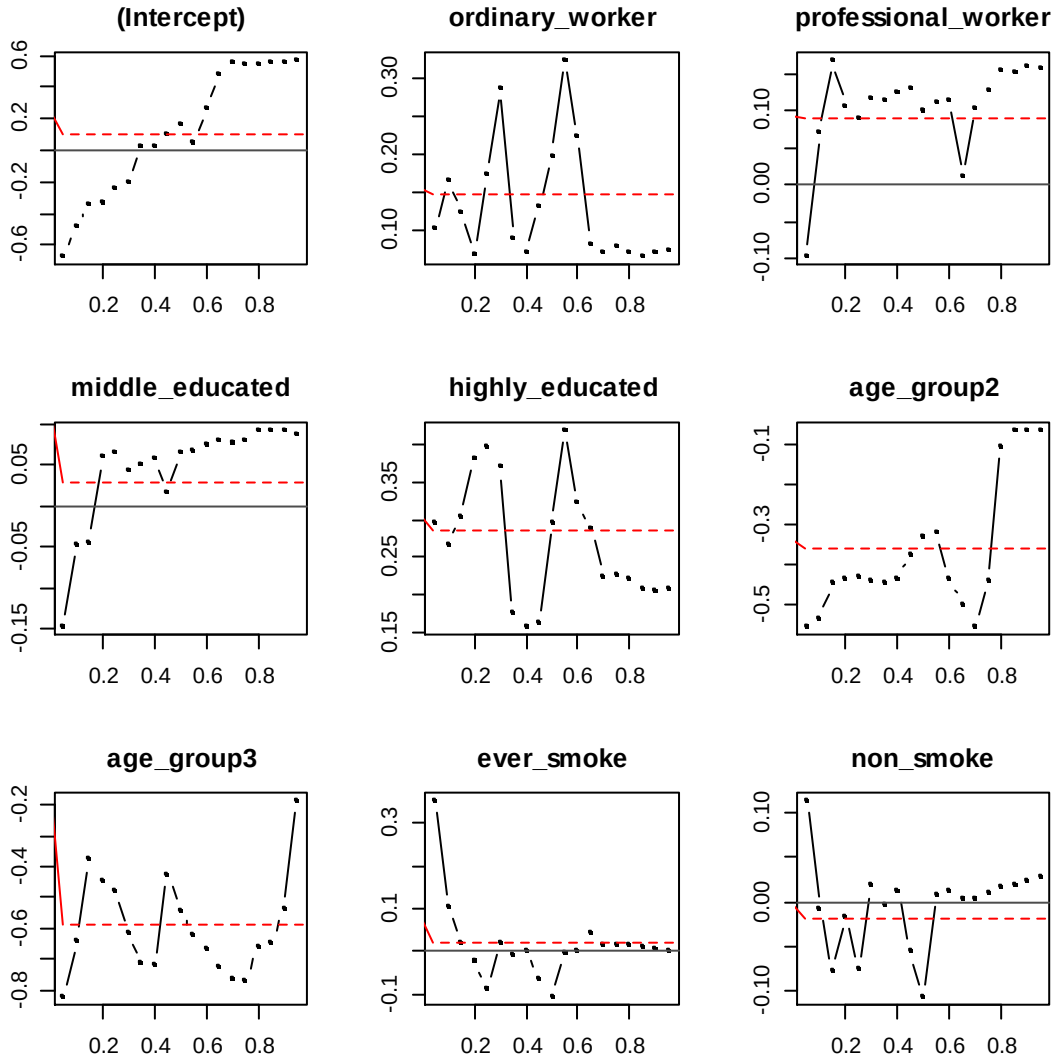

Figure 2. Multivariate Quantile Regression Estimates for Health Status Model.

Then, in order to identify the characteristic of healthy and unhealthy individual, we could summarize here that the health status distribution at the lower quantile indicate the unhealthy condition, meanwhile the health status in upper quantile reveal the healthy individual. Based on Table 3, we could determine that the characteristics of healthy individuals are indicated by significant independent variables at the upper quantile, there are ordinary and professional worker, midle and highly educated, young and non-smoking. Meanwhile, the characteristics of unhealthy individuals are indicated by significant independent variables at the lower quantile, there are highly educated and older than 30 years old.

V. DISCUSSION

The purposes of this present study are to identify the affecting of some independent variables over various quantile position of health status distribution and to determine the characteristics of healthy and unhealthy individual who are living at Hulu Langat, Malaysia. Based on some literatures, in the hypothesis model is assumed that health status could be affected by employment level, education level, age-group, smoking habits, physical exercise, average working and sleeping in a day, stress level, happiness in life and social problems that she/he has. This study uses quantile regression approach to evaluate the significant factors that affecting an individual's health status since quantile regression can assess the influences of each independent variable at various quantiles (including lower, middle and upper quantiles) of the dependent variable's distribution.

The NHMS III data which conducted at the year 2006 in Hulu Langat, a district at Malaysia is used in this study. After adjusting the hypothesis model with quantile regression technique to 530 respondents , this study found that at the middle quantile or at the median position, health status is determined significantly by employment level, education level, age-group and smoking habits. The factors that affecting the health status are different at the different quantiles. Since the distribution of health status at the lower quantile indicate unhealthy individual, thus this study yields that the characterictics of unhealthy individual are highly educated and older than 30 years old. Otherwise, the characteristics of healthy individual that located at the upper quantile of health status distribution are professional worker, highly educated, young (less than 31 years old) and non-smoking.

Thus, quantile regression provide more information on broader effects of socio-demography, life style and mental health characteristics on an individual's level of health status comparing to ordinary least square (OLS) regression. OLS approach assess the effect of different independent variables on dependent variable on central tendency (e.g. the mean) only. However, to assess factors affecting health status of an individual, effects on the mean of health status distribution are not the only concern. In the case of health status distribution, the mean represents a health status level in the middle value of health status, whereas the best condition of health status are contained at the upper range of health status distribution. In the contrary, the worst condition of an individual's health status is if the value of health status's distribution lying at the lower quantile.

ACKNOWLEDGEMENTS

The authors thank to the Institute for Public Health, Ministry of Health, Malaysia, 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. We also thank several anonymous referees for their constructive comments which have improved the final version of this paper.

REFERENCES

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

[2] Cheadle, A., Pearson, D., Wagner, E., Psaty, B.M., Diehr, P. & Koepsell, T. 1994. Relationship between socioeconomic status, health status, and lifestyle practices of American Indians: Evidence from a plains reservation population. *Public Health Report*, 109 (3): 405-413.

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

[4] Koenker, R. & Hallock, K.F. 2001. Quantile regression. *Journal of Prespectives* 15 (4): 143-156.

[5] Koenker, R. 2000. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. *Journal of Econometrics* 95: 347-374.

[6] 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

[7] Ramirez, M., Gallego, J.M. & Sepulveda, C.E. 2004. The determinants of the health status in a developing countries: results from the Colombian case. *Economia* 41:1-15.

[8] Shi, L. 1999. Sociodemographic characteristics and individual health behaviors. *Southern Medical Journal*, 91 (10): 933 – 941.

[9] Uitenbroek, D.G., Kerekovska, A. & Festchieva, N. 1996. Health lifestyle behavior and socio-demografic characteristics. A study of Varna, Glasgow and Edinburgh. *Social Science and Medicine* 43 (3): 367-377.

[10] Verzani, J. 2005. Using R for introductiory statistics. New York: Chapman & Hall/CRC Press Company.

[11] Yanuar, F., Ibrahim, K. & Jemain, A.A. 2010. On an application of structural equation for modeling of health index. *Environmental Health and Preventive Medicine*, 15: 285-291.

[12] Yanuar, F., Ibrahim, K. & Jemain, A.A. 2013. Bayesian Structural Equation Modeling For The Health Index. *Journal of Applied Statistics*, 40 (6): 1254-1269.

[13] Yu K. 2003. Quantile regression: Aplications and current research area. *The Statistician*, 52 (3): 331-350