KASU JOURNAL OF MATHEMATICAL SCIENCES (KJMS) VOL. 2, ISSUE 1, JUNE 2021
ISSN 2734-3839 (Print), ISSN 2735-962X (E-Copy)
http://www.journal.kasu.edu.ng/index.php/kjms

KASU JOURNAL OF MATHEMATICAL SCIENCES

Step Wise Approach to Structural Equation Modelling (SEM)

[1]Shehu S. U, [2]Sulaiman U., [3]Abdulazeez, A. S., [4]Shaib, I. O. and [5]Ahu, A. A.

[1,2,5]Department of Mathematics Statistics, Kaduna Polytechnic, Kaduna,

[3]Department of Mathematical Sciences, Kaduna State University, Kaduna

[4]Department of Statistics, Auchi Polytechnic, Nigeria.

e-mail: umar83@yahoo.com e-mail: shaibomade@gmail.com

Abstract

The trend of Structural Equation Modelling (SEM) applications in various disciplines is growing tremendously that the conceptual understanding of SEM becomes paramount interest of statisticians in the academics. Creating functional applications space for educating the learners and the teachers of statistics and mathematics the rudiments of SEM using the stepwise approach is a major focus of this work. In addition, this paper critical discussed the SEM fundamentals-specification, identification, Estimation discussed skeletally, SEM representation in Mulaik's convention and RAM diagram creation in R. It recommended further studies on the re-specification and modification of modelling in SEM.

Key words: SEM, RAM, Manifest and latent variables, Diagnostic check, Identification, Mulaik's.

1. Introduction

Structural Equation Modelling (SEM) is a general statistical modelling technique used to establish relationship among variables-measured and latent variables. Measured variables are independent variables that predict the dependent variables known as latent variables in SEM. For example it is often difficult to empirically measure the degree of wealth of an individual by merely asking the respondents. This may sound odd and intrusive to undertake statistically. However, investigation into this type of research questions can be possible through a painstakingly designed questionnaire in logical and coherence form to elicit information from the targeted respondents. This is usually the case in the social science research, on line phenomena and snow ball surveying. In most cases finding answers to question as to how wealthy individuals may require careful parameterisation involving measured variables or manifests and latent variables in the field of structural equation modelling. The research may attempt to test the relationship among measured variables; latent and measured variables; the extent of latent to latent variables and the nature of their relationships, these require complex regression modelling. In this paper, efforts are made to link confirmatory factor analysis (CFA) to SEM which is a robust method to address the aforementioned hypothetical statements.

1.2 Understanding SEM, Data Checking and Identification

SEM is an extended confirmatory factor analysis used to investigate models that are conceptually derived (apriori) to test if the theory fits the data set. SEM is a combination of factor analysis, CFA and multiple regressions modelling which is more of theory based than empirical (Shook, Ketchen, Hult and Kacmar2004). The followings have been identified as a family of SEM; Path Analysis, Path Modelling, Causal Modelling, Analysis of Variance and Covariance Structures, Latent Variable Analysis and Linear Structural Relation (LISREL) (Thelwall and Paul 2013). More importantly, the empirical and theoretical approaches to SEM cannot be fully implement without the knowledge of programming with specific software such as AMOS, LISREL, R and M-Plus. Each of this statistical software comes in versions. (Shook*et al* 2004) explained that the use of SEM needs diagnostic data procedure before it is considered suitable for application. It is also very vital for diagnostic test before SEM is applied because of problems resulting from outliers, normality, missing value and linearity. Therefore data checking becomes eminent in SEM. First, outlier data checking result in univariate and multivariate data sets. In univariate data for SEM, standardised Z-score is useful for data checking based on ordering using criteria scale point of -2.5 and +2.5 (Thelwall, Cugelman and Dawes2009). Any variable data set below or above the limit is considered outliers and causal factors need to be investigated before the SEM is applied. The same process is applicable to multivariate data set but uses Mahalanobis distance of linear regression technique based on the criteria. Normal distribution is the core of structural equation modelling. Bartholomew, Ntoumanis, Ryan, Bosch and Thøgersen (2011) suggested greater than or equal to 100 samples as normal for SEM. In univariate normal distribution, Q-Q plot of each variable against standard normal is plotted while multivariable plot Q-Q graph of all the variablessee graph in fig 1a.

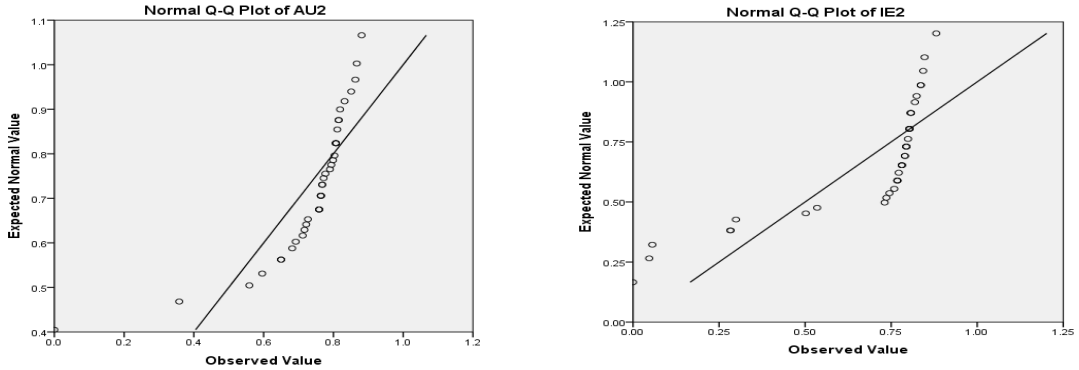


Figure 1a: Exploratory analysis of variable pattern Source: Minitab, 2020.

A deviation of data set from straight linear inform non-normality of data otherwise normal and fits the linear pattern or trend. However, variables that are not significantly nonnormal may be corrected by a transformation technique which is considered skewed then a log transformation is performed to produce a new variable data set which is approximately normally distributed (Shooket al 2004). Since non normal data provides strong evidence of unreliable analysis resulting

from data structure and pattern behaviour. Linear relationship existence in variables for SEM is necessary in data checking. Bartholomew etal (2011) indicated that relationship among all pairs of measured variables should be linear or random not following abnormal pattern that is deviation from linear form and proposed a potential linear fit. If it is non-linear pattern, there is need to transform the variable data sets to linear;If the variable possesses quadratic pattern then transform the variable taking the square root of one of the variables. If the relationship between two variables is multiplicative then logging or natural log transformation is appropriate to convert the variable into linear form. For instance, if z_1 is a product of variables x_1 and x_2 that is $z_1 = nx_1x_2$ then the transformation becomes $\log(z_1) = \log(n) + \log(x_1) + \log(x_2)$. This process is called deterministic law of nonlinear variable conversion. In addition, graphing for pattern detection is highly important to understand linearity behaviour of variable see the graph below in fig 1b.

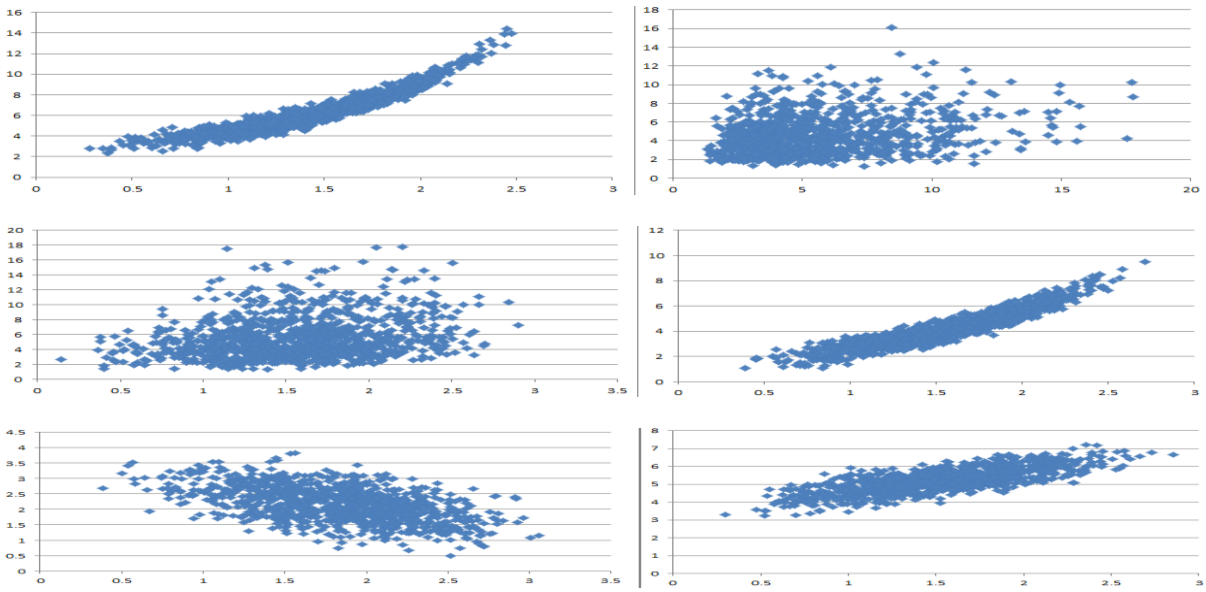


Figure 1b: Exploratory analysis of variable pattern Source: M-plus, 2020.

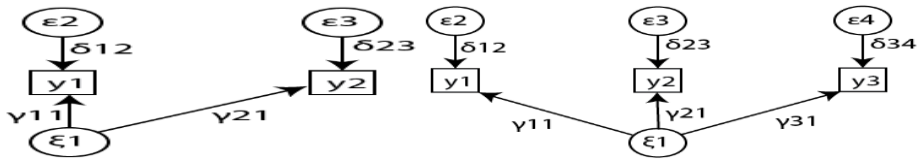
MacCullum, Browne and Sugawara(1996) pointed that the minimum sample size requirement for SEM to be adopted ranges from 100 and above or atleast 200 from small to medium samples while above 200 samples is considered as large samples. MacCullumet al(1996); Bartholomew, Ntoumanis, Ryan, andThøgersen (2011) considered relatively high subjects per parameter to be estimated in SEM. This indicates that if data is normally distributed then the effect size is expected to be large. This will give sufficient and strong evidence and chance of rejecting H_o. Small data does not enhance rejection of H_o which is the model fit based on Chi-squares test statistic.

ISSN 2734-3839 (Print), ISSN 2735-962X (E-Copy)
http://www.journal.kasu.edu.ng/index.php/kjms

2. Identification of SEM

There are various conditions in which SEM problem is not solvable or the solution is not meaningful because of properties of the raw data used to check it. Identification helps to check whether the solution of a model is appropriate for SEM procedure. The objective of identification is basically to assess whether the data set is consistent with a specified set of relationship between or among the variables as represented by a path diagram (MacCullum, *et al* (1996). Sometimes most diagrams cannot be tested because random data would be made to fit the model by the SEM algorithm. This is essentially the case when the model is too small for the number of variables in it. This situation therefore provides mathematical guide and a rule of thumb to ensure these issues do not arise in an analysis. Based on the path diagram, it is very easy for reader to describe SEM as under identify, just identify or over identify. This technique counts the number of measured, latent variable and parameters. The use of covariance is eminent. Consider the Path diagram and explanations in figure 2below:

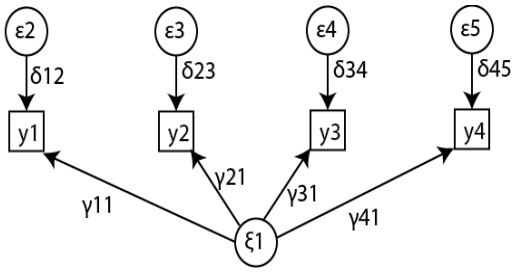
a. b.



$$y_1 = \gamma_{11}\xi_1 + \delta_{12}\varepsilon_2$$
$$y_2 = \gamma_{21}\xi_1 + \delta_{12}\varepsilon_3$$

$$y_1 = \gamma_{11}\xi_1 + \delta_{12}\varepsilon_2$$
$$y_2 = \gamma_{21}\xi_1 + \delta_{12}\varepsilon_3$$
$$y_2 = \gamma_{31}\xi_1 + \delta_{34}\varepsilon_4$$

c.



$$y_1 = \gamma_{11}\xi_1 + \delta_{12}\varepsilon_2$$
$$y_2 = \gamma_{21}\xi_1 + \delta_{12}\varepsilon_3$$
$$y_3 = \gamma_{31}\xi_1 + \delta_{34}\varepsilon_4$$
$$y_4 = \gamma_{41}\xi_1 + \delta_{45}\varepsilon_5$$

Consider the SEM in fig 2a. above simple (confirmatory factor analysis) model with one latent variable and 2 indicator variables and suppose that we want to use SEM to test the hypothesis that the indicator variable genuinely exists. This can be expressed in the diagram above, which is equivalent to the following two equations for the two exogenous variables. Recall that fitting an SEM model means selecting values of the free parameters so that the measured variances and covariances are as close as possible to the values predicted by the model. In this example, there are four free parameters $\gamma_{11}, \delta_{12}, \gamma_{21}, \delta_{23}$ and three measured variances and covariances,

$\sigma^2(y_1), \sigma^2(y_2), \sigma(y_1, y_2)$ from the sample data. Using the above equations, these can be related as follows:

$$\sigma^2(y_1) = \gamma^2{}_{11}\sigma^2(\xi_1) + \delta^2{}_{12}(\varepsilon_2)$$
$$\sigma^2(y_2) = \gamma^2{}_{21}\sigma^2(\xi_1) + \delta^2{}_{23}(\varepsilon_3)$$
$$\sigma(y_1, y_2) = \gamma_{11}\gamma_{21}\sigma^2(\xi_1)$$

In SEM it is normal to set the variances of all the latent variables to be 1. This is because the latent variables cannot be measured and their role is exerted in the model only through the paths in or out of them so their value is arbitrary. For example, if one latent variable variance was increased to 10 from 1 then multiplying the values of all path coefficients connected to it by 0.1 would result in an identical model (i.e., identical predicted measured variable variances and covariances). Hence, the above set of simultaneous equations can be reduced to the following.

$$\sigma^2(y_1) = \gamma^2{}_{11} + \delta^2{}_{12}$$
$$\sigma^2(y_2) = \gamma^2{}_{21} + \delta^2{}_{23}$$
$$\sigma(y_1, y_2) = \gamma_{11}\gamma_{21}$$

This is a set of three (non-linear) equations with four unknowns.So in theory, it could have infinitely many solutions and it does. This can be seen by rearranging the equations to express all the variables in terms of γ_{11} as follows (swapping the second and third equations).

$$\delta_{12} = \sqrt{\gamma^2{}_{11} - \sigma^2(y_1)},\ \gamma_{21} = \frac{\sigma(y_1, y_2)}{\gamma_{11}},\quad \delta_{23} = \sqrt{\sigma^2(y_2) - \gamma^2{}_{21}} = \sqrt{\sigma^2(y_2) - \left(\frac{\sigma(y_1, y_2)}{\gamma_{11}}\right)^2} \quad 1$$

Hence, given sample statistics there will typically be infinitely many choices of γ_{11} for which the above calculations yield values of the other free parameters $\delta_{12}, \gamma_{21}, \delta_{23}$ that make the model fit exactly (this doesn't work if is so that there are negative numbers inside the square roots). Hence (a) a model of the above form will always fit the data, no matter whether the model is correct or not, and (b) there are many choices of the parameters in the model and no way of choosing between them, it is ***under identified***.

Consider in fig. 2b another simple (confirmatory factor analysis) model with one latent variable and 3 indicator variables and suppose that we again want to use SEM to test the hypothesis that the indicator variable genuinely exists. This can be expressed in the diagram above, which is equivalent to the following three equations for the three exogenous variables.

In this example, there are six free parameters $\gamma_{11}, \delta_{12}, \gamma_{21}, \delta_{23}, \gamma_{31}, \delta_{34}$ and six possible measured variances and covariances

$$\sigma^2(y_1), \sigma^2(y_2), \sigma^2(y_3), \sigma(y_1, y_2), \sigma(y_2, y_3), \sigma(y_1, y_3).$$

These can be related as follows:Setting the variances of all the latent variables to 1, which we can always do, the above set of simultaneous equations can be reduced to the following.

$$\sigma^2(y_1)=\gamma^2{}_{11}\sigma^2(\xi_1)+\delta^2{}_{12}(\varepsilon_2) \quad \sigma^2(y_1)=\gamma^2{}_{11}+\delta^2{}_{12}$$
$$\sigma^2(y_2)=\gamma^2{}_{21}\sigma^2(\xi_1)+\delta^2{}_{23}(\varepsilon_3) \quad \sigma^2(y_2)=\gamma^2{}_{21}+\delta^2{}_{23}$$
$$\sigma^2(y_2)=\gamma^2{}_{31}\sigma^2(\xi_1)+\delta^2{}_{34}(\varepsilon_4) \quad \sigma^2(y_1)=\gamma^2{}_{11}+\delta^2{}_{12}$$
$$\sigma(y_1,y_2)=\gamma_{11}\gamma_{21}\sigma^2(\xi_1) \qquad \sigma(y_1,y_2)=\gamma_{11}\gamma_{21}$$
$$\sigma(y_1,y_3)=\gamma_{11}\gamma_{31}\sigma^2(\xi_1) \qquad \sigma(y_1,y_3)=\gamma_{11}\gamma_{31}$$
$$\sigma(y_2,y_3)=\gamma_{11}\gamma_{23}\sigma^2(\xi_1) \qquad \sigma(y_2,y_3)=\gamma_{11}\gamma_{23}$$

This is a set of six equations with six unknowns so in theory it could have one unique solution and it normally does (subject to the deltas and γ_{11}being non-negative anyway).Assuming that none of the square roots give imaginary numbers, the solution is as follows:

$$\gamma_{11}=\sqrt{\frac{\sigma(y_1,y_2)\sigma(y_1,y_3)}{\sigma(y_2,y_3)}},\gamma_{21}=\frac{\sigma(y_1,y_2)}{\gamma_{11}},\gamma_{31}=\frac{\sigma(y_1,y_3)}{\gamma_{11}},\delta_{12}=\sqrt{\sigma^2(y_1)-\gamma_{11}{}^2},$$
$$\delta_{23}=\sqrt{\sigma^2(y_2)-\gamma_{21}{}^2},\delta_{34}=\sqrt{\sigma^2(y_3)-\gamma_{31}{}^2}$$

Hence, given sample statistics there will typically be a solution, i.e., a choice of the free parameters $\gamma_{11},\delta_{12},\gamma_{21},\delta_{23},\gamma_{31},\delta_{34}$ that makes the model fit exactly. Hence (a) a model of the above form will almost always fit the data, no matter whether the model is correct or not, and (b) there is essentially one choice of the parameters in the model. In situations like this, if there is a unique solution for the free parameters to yield a perfect solution, the model is called ***just identified***. Normally a just identified model is useless for SEM purposes because it is impossible to test whether the model fits the data – since it will fit almost any data so there isnot enough evidence to make a test. A just-identified model can sometimes be used to test whether a specific parameter is non-zero but this is more common in path analysis than in SEM. In figure 2c, there are eight free parameters

$$\gamma_{11},\delta_{12},\gamma_{21},\delta_{23},\gamma_{31},\delta_{34},\gamma_{41},\delta_{45}$$

and ten possible measured variances

$$\sigma^2(y_1),\sigma^2(y_2),\sigma^2(y_3),\sigma^2(y_4)$$

and covariances,

$\sigma(y_1,y_2),\sigma(y_2,y_3),\sigma(y_1,y_3),\sigma(y_1,y_4),\sigma(y_2,y_4),\sigma(y_3,y_4)$.

These can be related as follow; setting the variances of all the latent variables to 1, which we can always do and this gives:

$$\sigma^2(y_1)=\gamma^2{}_{11}\sigma^2(\xi_1)+\delta^2{}_{12}(\varepsilon_2) \quad \sigma^2(y_1)=\gamma^2{}_{11}+\delta^2{}_{12}$$
$$\sigma^2(y_2)=\gamma^2{}_{21}\sigma^2(\xi_1)+\delta^2{}_{23}(\varepsilon_3) \quad \sigma^2(y_2)=\gamma^2{}_{21}+\delta^2{}_{23}$$
$$\sigma^2(y_3)=\gamma^2{}_{31}\sigma^2(\xi_1)+\delta^2{}_{34}(\varepsilon_4) \quad \sigma^2(y_3)=\gamma^2{}_{31}+\delta^2{}_{34}$$
$$\sigma^2(y_4)=\gamma^2{}_{41}\sigma^2(\xi_1)+\delta^2{}_{45}(\varepsilon_5) \quad \sigma^2(y_4)=\gamma^2{}_{41}+\delta^2{}_{45}$$
$$\sigma(y_1,y_2)=\gamma_{11}\gamma_{21}\sigma^2(\xi_1) \qquad \sigma(y_1,y_2)=\gamma_{11}\gamma_{21}$$
$$\sigma(y_1,y_3)=\gamma_{11}\gamma_{31}\sigma^2(\xi_1) \qquad \sigma(y_1,y_3)=\gamma_{11}\gamma_{31}$$
$$\sigma(y_2,y_3)=\gamma_{21}\gamma_{31}\sigma^2(\xi_1) \qquad \sigma(y_2,y_3)=\gamma_{21}\gamma_{31}$$
$$\sigma(y_1,y_4)=\gamma_{11}\gamma_{41}\sigma^2(\xi_1) \qquad \sigma(y_1,y_4)=\gamma_{11}\gamma_{41}\sigma^2(\xi_1)$$
$$\sigma(y_2,y_4)=\gamma_{21}\gamma_{41}\sigma^2(\xi_1) \qquad \sigma(y_2,y_4)=\gamma_{21}\gamma_{41}\sigma^2(\xi_1)$$
$$\sigma(y_3,y_4)=\gamma_{31}\gamma_{41}\sigma^2(\xi_1) \qquad \sigma(y_3,y_4)=\gamma_{31}\gamma_{41}$$

This is a set of ten equations with eight unknowns so, unless the fixed parameters; $\sigma^2(y_1), \sigma^2(y_2), \sigma^2(y_3), \sigma^2(y_4), \sigma(y_1,y_2), \sigma(y_2,y_3), \sigma(y_1,y_3), \sigma(y_1,y_4), \sigma(y_2,y_4), \sigma(y_3,y_4)$ are particularly nice, there will be no solution.

From a different perspective, if one of the four equations is removed then the model has one latent variable and 3 indicator variables and falls into the Example 2 case of a just identified model with a single unique solution. If equation 4 is removed then we would get one unique solution and if equation 3 is removed then we would get another unique solution, etc., giving four different candidate solutions to the full Example 3 problem. It would be a huge coincidence if all four different solutions were identical and this is very unlikely to happen in practice. Nevertheless, if the hypotheses underlying the model are true, then the solutions should be close. Hence the fact that we cannot get a solution to the problem allows us to test the hypotheses underlying the model by seeing how close we can get the model variances and covariances to the sample variances and covariances for the measured variables. In situations like this, where there is not a unique solution, a model is called **over identified**. An over identified model is necessary for testing hypotheses about structure in an SEM problem (i.e., the standard type of SEM problem).

2.1 The rule of Thumb for Identification

Most a time the rule of thumb for identification is most preferred to mathematical variance guide above. The formulae for identification based on the path diagram is expressed mathematically;

$df_M = \frac{n(n+1)}{2} - k$ where k is number of free parameter in a model (that is all the path coefficients and covariances between manifest variables), n is the manifest (measured) variable computed as $\frac{n(n+1)}{2}$. Where $k \le \frac{n(n+1)}{2}$, the SEM is over identified. The conditions for k when counted parameter for k excludes the following: (a) latent variables variance is usually set to 1 (b) Disturbance path coefficients (deltas) then k is computed by counting up all measured variable

using the $df_M = \frac{n(n+1)}{2} - k$. Therefore, the identification criteria are: if df_M is less than zero that $df_M < 0$, the SEM is under identified; if df_M equals zero that $df_M = 0$, the SEM is just identified and where df_M is greater than zero that $df_M > 0$, the SEM is over identified (Rigdon, 1994).

2.2 Testing Structural Equation Modelling

Testing SEM requires data and theory associated with the data. Hypothetical statements about data set in the following form: which are measured variable in the data? Which latent variable exist? Which latent variables relate to which other latent variables in the data? Therefore how are they related? Often times asked. These inform the basic three hypotheses to be tested in SEM which may be extended to sub hypothesis testing. To test a structural equation model, programme like LISREL, R with sem package and AMOS with SPSS optional package may be used. The test in SEM is the same like that of CFA. The data is feed into the package directly or indirectly by entering the data set or importing the data from excel.csv file or Notepad format. Data entering is often tedious and time consuming so the latter idea is considered appropriate for particularly larger sample. The diagram is drawn using AMOS tool or path programming in R sem package (Rigdon 1994). Again the null hypothesis is that the model fits the data and it is tested by the statistical programme fitting of SEM using data testing data and testing how close the fitted model is to the data. Bearing in mind the normality condition that the sample size that is prescribed by McCullem *et al* (1994) and Bartholomew *et al*(2011). This will enhance the test enough power to reject the model if it is a power fit. Standard goodness of fit test for structural equation model is Chi-squares test statistic. It is uniformly applicable in R, LISREL, M-plus and AMOS statistical software (Fox 2006). The probability value (P-value) associated with the Chi-square value and the degree of freedom can be used for significant test. Decision rule is; if P-value is less than the critical values (0.001, 0.01 and 0.05), it suggests poor model fit otherwise better model fit. However, a successful test would not prove the model is correct because another model may also fit the data reasonably well. This is because of theory behind the model. To address this, a more generally acceptable index is used to assess SEM.

2.2.1 RMSEA Index

Root Mean Square Error of Approximation Index (RMSEA-I) is a commonly used index of how well a model fit the data. It is an acceptable goodness of fit heuristic. In RMSEA index lower values indicate excellent or good model fit while higher values suggest poor or mediocre model fit. See the numerical definitive guide line RMSEA-I and decision (0.00-0.01, 0.01-0.05, 0.05-0.08, 0.08-0.99 and 0.1+ reveal excellent, good, mediocre, does not fit model and complete model fit rejection respectively (Fox 2006).

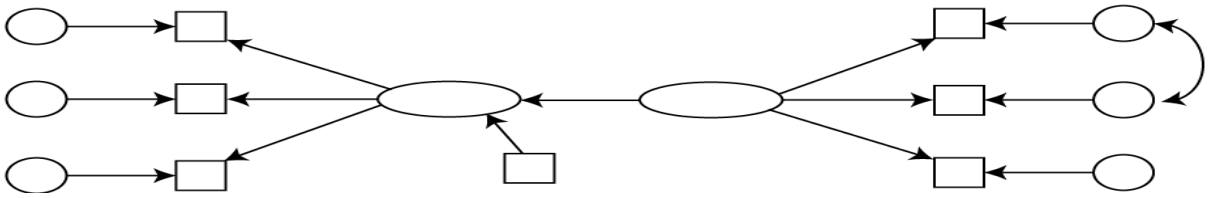
2.2.1.1 Interpretation of Chi-square result RMSEA-I and RMSEA-CI

The chi-square interpretation of SEM assessment is based on P-value; if P<0.05 then there is some evidence that the parameter is non-zero otherwise it is zero. The interest is the model fit the data

which is P>0.05 hence the parameter is zero. If the RMSEA-I falls within the 0.00-0.01 or 0.01-0.05, it indicates excellent or good model fit otherwise it is mediocre or poor. And if RMSEA-CI that is RMSEA Confidence Interval falls within an estimate value of 0.00 to 0.05 it indicate excellent or good fit.

3. SEM Representation

Path diagram and RAM diagram are the most useful representation of SEM. To construct a path diagram; two things are needed; a set of variables (measured or manifest/ unmeasured or latent) and set of hypothesised relations. Measured or manifest variables are represented by rectangles or squares and latent variables are represented by circles or ovals shape and arrow represent how variables are related. RAM diagram is a systematic way of representing networks in a simple text format in which each line is a comment containing: *relationship, variable name, variable value*. A causal connection is an arrow indicating source variable which is a cause of target variable. Any change may affect them simultaneously hence the need to study SEM notation arises (Stanley 2009; Fox 2006;Rigdon 1994 and McCullem *et al* 1994).By illustration see the examples below in figure 3.



3.1 Notation for SEM Diagram

SEM diagram comprises variables and differentiate between the two types irrespective of whether they are latent or manifest variable. Endogenous variables are those variables that are predicted by the model represented by the path diagram. In other hand, exogenous variables are not predicted by the model but reflect external input into the model. Generally in path diagram, endogenous variables are variables with at least one directed arrow pointing to them. Others are exogenous variables.

In SEM diagram naming convention particularly recommended to apply to variable and path coefficient. This concept helps to formulate SEM equations in a standard form. The adoption of Stanley (2009) Mulaik's notation for SEM diagram will be used to represent SEM equation.

3.1.1The Variable name Convention

This states that let notation x be a manifest exogenous variable (though rarely), y be a manifest endogenous variable, ε be a latent exogenous disturbance variable, η be a latent endogenous variable, ξ be a latent exogenous variable other than a disturbance, δ be a structural path coefficient for an arrow from disturbance to its endogenous variable γ a structural path coefficient

for a covariance between variables which is only allowed for exogenous variable (Stanley 2009; Thelwall*et al* 2013).

3.1.2 The Numbering Convention

Numbering convention according to Stanley (2009) deals with the variables and coefficient numbering pattern in SEM network path diagram, going by the guidelines: (1) Endogenous variables are labelled in one sequence (all variables η, then variable y in order of increasing numbers). (2) Endogenous variables are labelled in one sequence (all variables ξ, x, ε in order of increasing numbers) and none of these variables notations are the same in numbers. (3) Each structural coefficient has a double subscript as in $\gamma_{t,p}$ or $\delta_{s,p}$. The first subscript is the number of the arrow target and the second is the number of the arrow source. However, for convariances, the subscripts numbering could be in any order.
By illustration:

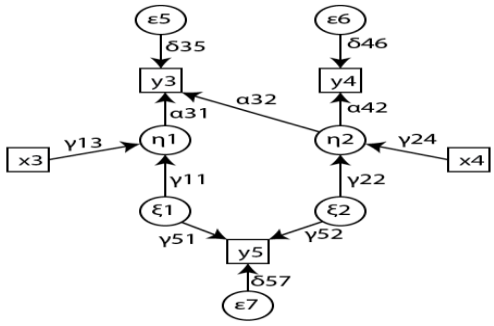
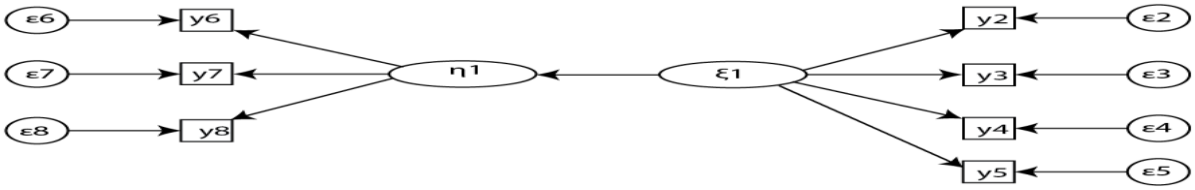


Figure 4: Label the variables diagram using Mulaik's rules.

This is suggested in Fox (2006) Causal Linear Model that Mulaik's notation create a standard for RAM diagram specification for SEM. When creating a RAM path diagrams in SEM package, there is a modification to the above format naming convention. (1) the disturbances are removed and replaced with residual variance variables for the associated variables and the residual variance of latent variables are set to 1.

4. Implementing RAM diagram in R

The diagram below represent the convention of Mulaik's notation and creating RAM for SEM specificatiom model (Fox 2006) therefore the network path diagram below transformed indicates the Mulaik' notation for SEM.



ISSN 2734-3839 (Print), ISSN 2735-962X (E-Copy)
http://www.journal.kasu.edu.ng/index.php/kjms

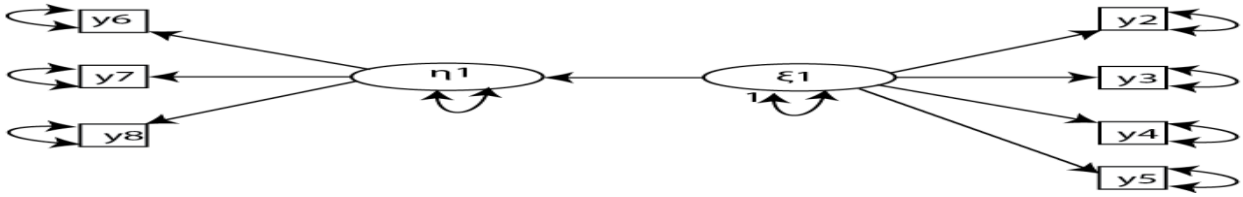


Figure 5: Reduction variables diagram using Mulaik's rules.

Its path diagram file is thus:
##Relationships (variables are factor loadings here)

E1 -> y2, E1y2, NA

E1 -> y3, E1y3, NA

E1 -> y4, E1y4, NA

E1 -> y5, E1y5, NA

n1 -> y6, n1y6, NA

n1 -> y7, n1y7, NA

n1 -> y8, n1y8, NA

E1 -> n1, E1n1, NA

##Measured variables (variables are residual variances here)

y2 <-> y2, y2ResVar, NA

y3 <-> y3, y3ResVar, NA

y4 <-> y4, y4ResVar, NA

y5 <-> y5, y5ResVar, NA

y6 <-> y6, y6ResVar, NA

y7 <-> y7, y7ResVar, NA

y8 <-> y8, y8ResVar, NA

##Latent (factor) variables (variables are variances or residual variances)

n1 <-> n1, n1ResVar, 1

E1 <-> E1, E1ResVar, 1

##

5. Estimation Procedure

In the early part of this paper, some estimation procedures are identified. Variances and Covariances are at the heart of SEM and they are used to calculate and predict the variances of

exogenous variables. The accuracy of SEM depends on the extent to which the Variances and Covariances exogenous variables. Estimations such as model fit, assessment, strength, correlation. Validity and relationships are performed using chi-squares, RMSEA-I, CI, variances and covariances of multivariate procedures and multiple regression modelling. Average Variance Extraction (AVE) is used to test the convergent validity, composite reliability and discriminant validity using factor analysis and principal component analysis. Matrix notation and selection is most appropriate for manual computational procedure for SEM problems.

5.1 Matrix Notation and Selection Alternative Techniques

The basic column vectors of different types of variables are:

$\boldsymbol{y} = \begin{bmatrix} y_3 \\ y_4 \\ y_5 \end{bmatrix}, \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$. Combining: $\boldsymbol{\eta}^* = \begin{bmatrix} \eta_1 \\ \eta_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}, \boldsymbol{\xi}^* = \begin{bmatrix} \xi_1 \\ \xi_2 \\ x_3 \\ x_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$.

The matrix A must have 5 rows and 5 columns since it multiplies the 5 endogenous variables to give factor for the 5 endogenous variables. Similarly, the matrix $\boldsymbol{\Gamma}$ must have 7 rows and 5 columns since it multiplies 7 exogenous variables to give factors for the 7 endogenous variables.

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \alpha_{31} & \alpha_{32} & 0 & 0 & 0 \\ 0 & \alpha_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{and} \boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & 0 & \gamma_{13} & 0 & 0 & 0 & 0 \\ 0 & \gamma_{22} & 0 & \gamma_{24} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_{35} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta_{46} & 0 \\ \gamma_{51} & \gamma_{52} & 0 & 0 & 0 & 0 & \delta_{57} \end{bmatrix}$$

Thus: $\boldsymbol{\eta}^* = \boldsymbol{A}\boldsymbol{\eta}^* + \boldsymbol{\Gamma}\boldsymbol{\xi}^*$becomes:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \alpha_{31}\eta_1 + \alpha_{32}\eta_2 \\ \alpha_{42}\eta_2 \\ 0 \end{bmatrix} + \begin{bmatrix} \gamma_{11}\xi_1 + \gamma_{13}x_3 \\ \gamma_{22}\xi_2 + \gamma_{24}x_4 \\ \delta_{35}\varepsilon_5 \\ \delta_{46}\varepsilon_6 \\ \gamma_{51}\xi_1 + \gamma_{52}\xi_2 + \delta_{57}\varepsilon_7 \end{bmatrix} = \begin{bmatrix} \gamma_{11}\xi_1 + \gamma_{13}x_3 \\ \gamma_{22}\xi_2 + \gamma_{24}x_4 \\ \alpha_{31}\eta_1 + \alpha_{32}\eta_2 + \delta_{35}\varepsilon_5 \\ \alpha_{42}\eta_2 + \delta_{46}\varepsilon_6 \\ \gamma_{51}\xi_1 + \gamma_{52}\xi_2 + \delta_{57}\varepsilon_7 \end{bmatrix}$$

5.2 Alternative forms of the SEM matrix equation

One problem with the standard matrix form of the SEM equations is that there are two copies of the same vector, which makes algebra more difficult with it. We can easily resolve this issue with some matrix algebra, however.$\boldsymbol{\eta}^* = \boldsymbol{A}\boldsymbol{\eta}^* + \boldsymbol{\Gamma}\boldsymbol{\xi}^*$the standard form, $\boldsymbol{\eta}^* - \boldsymbol{A}\boldsymbol{\eta}^* = \boldsymbol{\Gamma}\boldsymbol{\xi}^*$ - subtracting from both sides, $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\eta}^* = \boldsymbol{\Gamma}\boldsymbol{\xi}^*$ - factorising (non-commutative), $\boldsymbol{\eta}^* = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}^*$ - multiplying both sides on the left by $(\boldsymbol{I} - \boldsymbol{A})$. In this form of the equation, no variables occur twice. Note that the last stage will only work if $(\boldsymbol{I} - \boldsymbol{A})$ has an inverse. For convenience, sometimes we replace $(\boldsymbol{I} - \boldsymbol{A})$ by $\boldsymbol{B}$, giving the reduced form:$\boldsymbol{\eta}^* = \boldsymbol{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}^*$.

Note that in this form of the equation, the endogenous variables are affected only by the exogenous variables. This is possible because the matrix inversion process creates a matrix that includes the effect of the endogenous to endogenous direct links by redirecting them through related exogenous links.

5.3 An equation for the manifest variables

The variance and covariance calculations that can be made from the raw data gathered for the variables in a SEM are those between the manifest variables **x** and **y**. To help represent the calculations needed, another matrix can be introduced.

$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix}$ is a column vector of all the manifest variables.

$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{\eta}^* \\ \boldsymbol{\xi}^* \end{bmatrix}$ is a column vector of all the variables (endogenous at the top, exogenous at the bottom).Note that $\boldsymbol{Z} = \boldsymbol{Gv}$ since the selector matrix $\boldsymbol{G}$ merely selects the manifest variables from the complete set of variables. Hence:$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \boldsymbol{Gv} = \begin{bmatrix} \boldsymbol{G_y} & 0 \\ 0 & \boldsymbol{G_x} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}^* \\ \boldsymbol{\xi}^* \end{bmatrix}$ But using the reduced form:

$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{G_y} & 0 \\ 0 & \boldsymbol{G_x} \end{bmatrix} \begin{bmatrix} \boldsymbol{B^{-1}\Gamma\xi^*} \\ \boldsymbol{\xi}^* \end{bmatrix}$, $Z = \begin{bmatrix} \boldsymbol{G_y} & 0 \\ 0 & \boldsymbol{G_x} \end{bmatrix} \begin{bmatrix} \boldsymbol{B^{-1}} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Gamma\xi^*} \\ \boldsymbol{\xi}^* \end{bmatrix}$,$Z = \begin{bmatrix} \boldsymbol{G_y} & 0 \\ 0 & \boldsymbol{G_x} \end{bmatrix} \begin{bmatrix} \boldsymbol{B^{-1}} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Gamma} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{\xi}^*$

And if we define new matrices$\boldsymbol{B}^* = \begin{bmatrix} \boldsymbol{B} & 0 \\ 0 & \boldsymbol{I}_{|\xi^*|} \end{bmatrix}$ and $\boldsymbol{\Gamma}^* = \begin{bmatrix} \boldsymbol{\Gamma} \\ \boldsymbol{I}_{|\xi^*|} \end{bmatrix}$ (unless there are no x variables, in which case $\boldsymbol{B}^* = \boldsymbol{B}$and $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma}$**)** then this equation can be written as. $\boldsymbol{Z} = \boldsymbol{GB^{*-1}\Gamma^*\xi^*}$

5.4 Variance and covariance calculations

Assuming that the variables are normalised (or at least have the mean subtracted) then the sum of squares of the data is the variance and the variance-covariance matrix $\boldsymbol{\Sigma}$ of the manifest variables is given by $\boldsymbol{\Sigma = ZZ^T = GB^{*-1}\Gamma^*\xi^*(GB^{*-1}\Gamma^*\xi^*)^T}$, $\boldsymbol{\Sigma = GB^{*-1}\Gamma^*\xi^*\xi^{*T}\Gamma^{*T}B^{*T-1}G^T}$

Now the middle part of the system is the variance covariance matrix for the exogenous variables $\boldsymbol{\xi^*\xi^{*T}}$and we denote this by $\boldsymbol{\Phi}$(phi). Here $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi_{\xi\xi}} & \boldsymbol{\Phi_{\xi x}} & \boldsymbol{0} \\ \boldsymbol{\Phi_{x\xi}} & \boldsymbol{\Phi_{xx}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Phi_{\varepsilon\varepsilon}} \end{bmatrix}$

Where $\boldsymbol{\Phi_{\xi\xi}}$ is the variance-covariance matrix for the exogenous latent variables, $\boldsymbol{\Phi_{xx}}$ is the variance- covariance matrix for the manifest exogenous variables, $\boldsymbol{\Phi_{\xi x}}$ is the covariance matrix for the latent and manifest exogenous variables, and $\boldsymbol{\Phi_{\varepsilon\varepsilon}}$ is the variance-covariance matrix for the disturbances. Hence the variance-covariance matrix of the manifest variables $\boldsymbol{\Sigma}$ can be written in an alternative form as$\boldsymbol{\Sigma = GB^{*-1}\Gamma^*\Phi\Gamma^{*T}B^{*T-1}G^T}$**.**

6. Application Area of SEM

In line with the motivation of this paper, SEM has many applications in various walks of life ranging from social science, earth sciences- mining and geology, geography; sociology, psychology research into intelligence and personality, politics, economic modelling, finance,

accident research, environmental, taxonomy, biological sciences, medicine, religious research and widely used today is in information technology and web metrics analysis. The sources for structural equation are acquired through primary source of data especially the use of coherent, sequential and logically integrated questionnaires. Sometimes observation and experimentation may be thoughtful (Thelwall and Paul 2013; Thelwall and *et al* 2009; Fox 2006; Richard *et al* 1988).

7. Conclusion and Recommendation

Stepwise structural equation model theories, principles and applications have been discussed somewhat extensively and basic techniques explained. The vague idea of SEM has also been explicit and the major steps in SEM such as specification, identification, Estimation were discussed. SEM representation in Mulaik's convention and RAM diagram creation in R have been treated very clearly. However, the paper does not give clue to the re-specification/ modification of model and real life data application. Therefore recommends advance learning and understanding of SEM by researchers in the field of multivariate and other related areas.

References

Bartholomew, K. J., Ntoumanis, N., Ryan, R. M., Bosch, J. A., and Thøgersen Ntoumani, C. (2011). Self-determination theory and diminished functioning:the role of interpersonal control and psychological need thwarting. Personality and Social Psychology Bulletin, 37, 1459 e1503.

Bartholomew, K. J., Ntoumanis, N., Ryan, R. M., andThøgersen-Ntoumani, C. (2011). Psychological need thwarting in the sport context: assessing the darker side ofathletic experience. *Journal of Sport and Exercise Psychology*, 33, 75e102.

Fox (2006). Structural equation modeling with the sem Package in R; Structural Equation Modeling, 13:465-486. Available at http://socserv.mcmaster.ca/jfox/Misc/sem/SEM-paper.pdf

Richard P. Bagozzi, Youjae Yi (1988), On the evaluation of structural equation models, *Journal of the Academy of Marketing Science*, Vol. 16, No. 1. pp. 74-94.

Rigdon, E.E. (1994) Calculating degrees of freedom for a structural equation model.*Structural Equation Modeling*, *1*, 274-278.

Shook C.L., Ketchen J.D., Hult G.T.M., and Kacmar K.M. (2004) An assessment of the use of structural equation modelling in strategic management research. *Strategic Management Journal*, 25, 397-404.

Stanley A. Mulaik (2009) "Linear casual Modeling with structural equations" Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series, USA, Taylor and Francis Group

Thelwall, M., Cugelman, B., & Dawes, P. (2009)The dimensions of website credibility and their relation to active trust and behavioural impact, Communications of the Association for Information Systems, 24, 455-472.

Thelwall, M., and Paul, W.(2013) Advanced statistical theory and modelling, Concept and Theory, University of Wolverhampton, *Journal of Statistical Education*,[Online] Vol. 46, No. 23. pp. 14-32 [Accessed 25 July 2013]. Available at: http://www.amstat.org/publications/jse/jse_data_archive.htm.