CME Article

Biostatistics 308. Structural equation modeling

Y H Chan

Yong Loo Lin School
of Medicine
National University
of Singapore
Block MD11
Clinical Research
Centre #02-02
10 Medical Drive
Singapore 117597

Y H Chan, PhD
Head
Biostatistics Unit

Correspondence to:
Dr Y H Chan
Tel: (65) 6874 3698
Fax: (65) 6778 5743
Email: medcyh@
nus.edu.sg

The techniques discussed in our series, thus far, examine unidirectional relationships – i.e. how the independent variables affect the dependent variable. The assumptions were that the dependent response is random and subject to error whereas the independent variables could be measured directly (error-free), interdependency or simultaneous causation among these independent variables were not modelled. Multicolinearity among the independent variables is an issue which we could resolve using PCA or Factor analysis[2] to derive independent components/ factors for modelling purposes, given that meaningful interpretations are feasible.

Structural equation model (SEM) is used to examine multiple and interrelated dependence relationships and able to take into account the measurement error of the independent variables. The aim of this article (the finale of our series) is to introduce the basic concepts of this dynamically growing technique. SEM has other common names (just to mention a few) – covariance structure analysis, latent variable analysis, LISREL analysis, causal modeling, path analysis, dependence analysis, confirmatory factor analysis. Commonly-used SEM software includes LISREL, AMOS, EQS and SAS CALIS. The syntax and outputs of each of the software are different – we will discuss the use of SAS CALIS (covariance analysis of linear structural equations) in this article.

Consider a hypothetical example where one collects the overall STRESS level of a subject and his scores on "poor" personal HEALTH, low FINANCE status, FAMILY unhappiness and WORK dissatisfaction. All scores are measured on a scale of 0 (low) to 100 (high), higher scores indicate higher stress, poorer health, lower finance level, more family unhappiness and greater work dissatisfaction. Table I shows the descriptive scores for the 350 surveyed subjects.

The usual analysis to determine which of the sub-scores significantly affect the overall stress score, a multiple linear regression[1] is performed (Table II and Fig. 1). All predictors were significant but the "direction" of Health and Finance with Stress is "opposite" (we know this is due to multicolinearity, there is a significantly high correlation between Work and Finance, Health and Family, see Table III).

Fig. 1 Regression model for Stress.

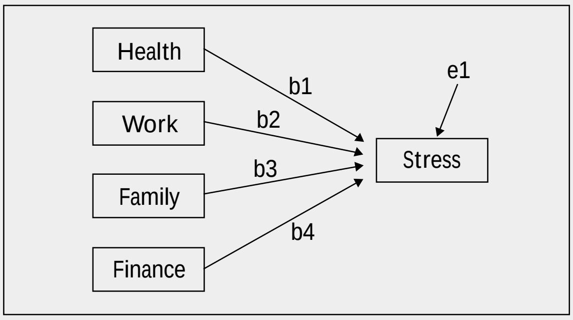


Table I. Descriptive statistics for the scores.

	Descriptive statistics				
	N	Minimum	Maximum	Mean	Std. deviation
STRESS	350	23	82	61.01	14.71
HEALTH	350	39	87	60.00	9.96
WORK	350	28	82	63.26	14.09
FAMILY	350	33	83	57.38	11.51
FINANCE	350	52	87	66.97	8.95
Valid N (listwise)	350				

NB: the Stress score is not the accumulated sum of the sub-scores.

Table II. Linear regression model for the overall Stress score.

Coefficients[a]

Model		Unstandardised coefficients		Standardised coefficients			95% Confidence interval for B	
		B	Std. error	Beta	t	Sig.	Lower bound	Upper bound
1	(Constant)	16.064	5.619		2.859	.005	5.012	27.116
	HEALTH	-.169	.073	-.115	-2.306	.022	-.314	-.025
	WORK	.519	.066	.497	7.902	.000	.390	.648
	FAMILY	.644	.068	.503	9.412	.000	.509	.778
	FINANCE	-.219	.099	-.133	-2.203	.028	-.415	-.023

[a]: Dependent variable: STRESS.

Table III. The correlations among the variables.

Coefficients[a]

		STRESS	HEALTH	WORK	FAMILY	FINANCE
STRESS	Pearson correlation	1.000	.257**	.528**	.573**	.239**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	350	350	350	350	350
HEALTH	Pearson correlation	.257**	1.000	.154**	.613**	.104
	Sig. (2-tailed)	.000	.	.004	.000	.053
	N	350	350	350	350	350
WORK	Pearson correlation	.528**	.154**	1.000	.292**	.738**
	Sig. (2-tailed)	.000	.004	.	.000	.000
	N	350	350	350	350	350
FAMILY	Pearson correlation	.573**	.613**	.292**	1.000	.035
	Sig. (2-tailed)	.000	.000	.000	.	.511
	N	350	350	350	350	350
FINANCE	Pearson correlation	.239**	.104	.738**	.035	1.000
	Sig. (2-tailed)	.000	.053	.000	.511	.
	N	350	350	350	350	350

**: Correlation is significant at the 0.01 level (2-tailed).

Let us introduce the flavour of using SEM.

The following SAS statements using Proc CALIS will reproduce the results in Table II.

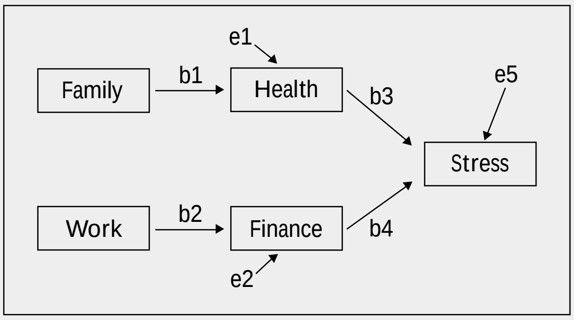
```
proc calis data = sem_eg cov;
lineqs Stress = b1 Health + b2 Work + b3 Family + b4
Finance + e1;
std
      e1 = ve1;
var Health Work Family Finance Stress;
run;
```

If the term "cov" is omitted, the default correlation matrix will be used and only the standardised coefficients (Beta) will be given. The parameters are the regression coefficients b1 to b4 and the variance ve1 of the error term e1 (convention: names of error terms must begin with the letter "e"). The error term e1 models the random error of Stress. There is no need to have an * between b1 and Health to indicate the multiplication of the variable by the coefficient. If the name of a coefficient (for example b1) is left out, the value of the coefficient is assumed to be 1.

The "std" section specifies the variances of the variables that are not error-free (convention: must begin with "v" or "var" followed by any alphabet – recommended to be the name of the error term for easy referencing). Leaving out the name of a variance, assumes that the variance is 0. The "var" statement (after the "std" section) is optional (recommended for efficient SAS computations). SEM could also be performed using the means, standard deviations and correlations of the data (instead of the raw data), see Appendix I for the SAS codes to input the estimates before running the SEM analysis.

SEM is a confirmatory type of analysis rather than exploratory. Let us postulate a possible structural equation model where Family is mediated by Health and Work is mediated by Finance (Fig. 2).

Fig. 2 Structural equation model: Family mediated by Health and Work mediated by Finance.



The SAS codes are:

```
proc calis data = sem_eg stderr;
lineqs  Health = b1 Family + e1,
        Finance = b2 Work + e2,
        Stress = b3 Health + b4 Finance + e5;
std
        Family = vf,
        Work = vw,
        e1 = ve1,
        e2 = ve2,
        e5 = ve5;
run;
```

Table IV shows the parameter estimates of the above model (Fig. 2).

The direction of the parameter estimates indicates the effects on Stress. More Family unhappiness indicates more poor Health which in turn affects Stress. Similarly, higher Work dissatisfaction indicates lower Finance status and more Stress. Statistical significance is achieved when the absolute of the z-value exceeds 1.96. Other information that we can gather is that Work and Family each explained nearly 7%, Health about 5%, Finance 3.5% and an unexplained variance of 7% on the outcome Stress.

Another possible application of SEM is to model the measurement error of an independent variable which is ignored during a linear regression analysis. The impact of this ignorance of the measurement error is an underestimation of our results since (observed = true X reliability of the estimate). Unless the reliability of the estimate is 100%, the observed relationship is always an underestimate of the true relationship.

Let us assume that there is some measurement error in Health.

The set of codes to model both the measurement error of Stress and Health:

```
proc calis data = sem_eg;
lineqs Stress = b1 fh + b2 Work + b3 Family + b4
          Finance + e1,
           Health = fh + e2;
std
          fh = vfh,
          e1 = ve1,
          e2 = ve2;
run;
```

The variable “‘h” is known as a latent variable (the “true” value of Health) which cannot be measured directly (convention: must begin with the letter “f”).

Table IV. Parameter estimates – SEM of Fig. 2.

Parameter	Variable	Estimate	SE	z-value
b1	Family → Health	0.6134	0.0423	14.51
b2	Work → Finance	0.7377	0.0361	20.41
b3	Health → Stress	0.2344	0.0509	4.61
b4	Finance → Stress	0.2151	0.0509	4.23
Variances				
vf	Family	1.0000	0.0695	14.39
vw	Work	1.0000	0.0695	14.39
ve1	Health	0.6237	0.0472	13.21
ve2	Finance	0.4558	0.0345	13.21
ve5	Stress – unexplained variance	0.8883	0.0673	13.21

e2 models the measurement error of Health. If one believes that there is minimal measurement error in the independent variables, SEM serves no additional advantage over linear regression.

The final application that we want to discuss is confirmatory factor analysis (CFA). Let us say, the variable Stress is not available and we believe that there are 2 domains of stress (Well-being and Needs) from the 4 variables collected. Performing a simple Factor analysis[(2)], using eigenvalues >1 criterion, shows that Well-being is made-up of Health and Family whereas Needs is from Work and Finance (Table V).

Table V. Simple factor analysis using eigenvalues >1 criterion.

Rotated component matrix[a]		
	Component	
	1	2
HEALTH	5.407E-02	.888
WORK	.917	187
FAMILY	9.807E-02	.900
FINANCE	.939	-1.809E-02

Extraction method: principal component analysis.

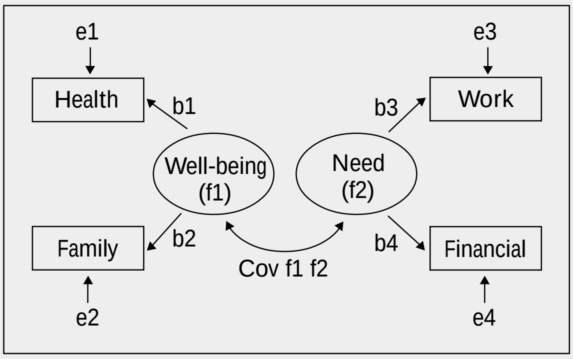
Rotation method: Varimax with Kaiser normalisation.

[a.]: Rotation converged in 3 iterations.

A CFA is performed to verify the above a priori model (Fig. 3). The covariance between Well-being and Needs (cov f1 f2) determines the relationship between the 2 latent variables. The SAS codes to model these 2 latent variables (Well-being and Needs) are:

```
Proc calis data = sem_eg;
lineqs  Health = b1 f1 + e1,
        Family = b2 f1 + e2,
        Work = b3 f2 + e3,
        Finance = b4 f2 + e4;
std
        e1 = ve1,
        e2 = ve2,
        e3 = ve3,
        e4 = ve4,
        f1 = vf1,
        f2 = vf2;
cov
        f1 f2 = covf1f2;
run;
```

Fig. 3 Structural equation model for confirmatory factor analysis.



High loadings (at least 0.7) in b1 & b2 on Well-being and b3 & b4 on Needs "confirms" the existence of the a priori Fig. 3 model.

There are several statistics (which are provided in the course of SAS CALIS analysis) for checking on the adequacy of the SEM (Table VI).

Table VI. Model fitting statistics for SEM.

Statistics	
1. Chi-square estimate	This indicates the amount of difference between the observed and expected covariance matrices. A chi-square value close to zero with p>0.05 indicates a good fit
2. Comparative fit index (CFI)	This is the discrepancy function adjusted for sample size. Ranges from 0 to 1, at least >0.9 for an acceptable model fit
3. Root mean square error of approximation (RMSEA)	This is related to the residual in the model. Ranges from 0 to 1, at least <0.06 for an acceptable model fit

We had just merely scraped the tip of the SEM iceberg. The examples discussed are the usual applications of SEM. A suggestive reading list is provided and do seek the help of a biostatistician in the event of involved relationships in a model. I hope you have enjoyed our whole series of Basic Statistics for Doctors. For completeness, the references for the whole series are given in Appendix II.

REFERENCES

1. Chan YH. Biostatistics 201. Linear regression analysis. Singapore Med J 2004; 45:55-61.
2. Chan YH. Biostatistics 302. Principal component and factor analysis. Singapore Med J 2004; 45:558-66.

SUGGESTED READING LIST

Bollen KA. Structural Equations with Latent Variables. New York: Wiley, 1989.

Eye AV. Latent Variables Analysis. Thousand Oaks, CA: Sage Publications, 1994.

Hoyle RH. Structural Equation Modeling: Concepts, Issues, and Applications. Thousand Oaks, CA: Sage Publications, 1995.

Kline RB. Principles and Practice of Structural Equation Modeling. New York: Guilford Press, 1998.

Loehlin JC. Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis. Mahwah, NJ: L Erlbaum Associates, 1998.

Maruyama GM. Basics of Structural Equation Modeling. Thousand Oaks, CA: Sage Publications, 1998.

Schumaker RE. A Beginner's Guide to Structural Equation Modeling. Mahwah, NJ: L Erlbaum Associates, 1996.

Appendix I. Using the mean, standard deviations and the correlation estimates (from Tables I & III) to run the above SEM analyses.

```
data sem_eg (type=corr);
input _type_ $1-4 name_ $6-12
               health 14-19 work 21-26 finance 28-33
               family 35-40 stress 42-47;
               cards;
n              350  350  350  350  350
mean           60.0  63.2  66.9  57.3 61.0
std            9.9  14.1   8.9  11.5 14.7
corr health    1.00
corr work      0.15  1.00
corr finance   0.10  0.74  1.00
corr family    0.61  0.29  0.04 1.00
corr stress    0.26  0.53  0.24 0.57  1.00
;;;;
```

Appendix II. Topics in the Singapore Medical Journal series on Basic Statistics for Doctors.

1. Randomised controlled trials (RCTs) – essentials. Singapore Med J 2003; 44:60-3.
2. Randomised controlled trials (RCTs) – sample size: the magic number Singapore Med J 2003; 44:172-4.
3. Biostatistics 101. Data presentation. Singapore Med J 2003; 44:280-5.
4. Biostatistics 102. Quantitative data – parametric and non-parametric tests. Singapore Med J 2003; 44:391-6.
5. Biostatistics 103. Qualitative data - tests of independence. Singapore Med J 2003; 44:498-503.
6. Biostatistics 104. Correlational analysis. Singapore Med J 2003; 44:614-9.
7. Biostatistics 201. Linear regression analysis. Singapore Med J 2004; 45:55-61.
8. Biostatistics 202. Logistic regression analysis. Singapore Med J 2004; 45:149-53.
9. Biostatistics 203. Survival analysis. Singapore Med J 2004; 45:249-56.
10. Biostatistics 301. Repeated measurement analysis. Singapore Med J 2004; 45:354-68.
11. Biostatistics 301a. Repeated measurement analysis (mixed models). Singapore Med J 2004; 45:456-60.
12. Biostatistics 302. Principal components and factor analysis. Singapore Med J 2004; 45:558-66.
13. Biostatistics 303. Discriminant analysis. Singapore Med J 2005; 46:54-61.
14. Biostatistics 304. Cluster analysis. Singapore Med J 2005; 46:153-9.
15. Biostatistics 305. Multinomial logistic regression. Singapore Med J 2005; 46:259-68.
16. Biostatistics 306. Log-linear models: poisson regression. Singapore Med J 2005; 46:377-87.
17. Biostatistics 307. Conjoint analysis and canonical correlation. Singapore Med J 2005; 46:514-17.
18. Biostatistics 308. Structural equation modeling. Singapore Med J 2005: 46:675-80.

EDITOR'S NOTE

This article on "Structural equation modeling" concludes the "Basic Statistics for Doctors" series of the Singapore Medical Journal (SMJ). The SMJ wishes to thank Dr Chan Yiong Huak for authoring all 18 articles in this well-received series over the past three years. Dr Chan will continue to serve on the SMJ Editorial Board as the resident biostatistics expert.

Professor Wilfred C G Peh
Editor
Singapore Medical Journal

SINGAPORE MEDICAL COUNCIL CATEGORY 3B CME PROGRAMME

Multiple Choice Questions (Code SMJ 200512A)

	True	False
Question 1. The following statistics are the goodness of fit measures for a structural equation model (SEM):		
(a) The p-values of the estimates.	□	□
(b) The comparative fit index (CFI).	□	□
(c) The root mean square error of approximation (RMSEA).	□	□
(d) The chi-square estimate.	□	□
Question 2. The following denotes a good fit for a SEM:		
(a) CFI < 0.9.	□	□
(b) A significant p-value with chi-square value near 0.	□	□
(c) RMSEA < 0.06.	□	□
(d) Significant p-values for the estimates.	□	□
Question 3. The SEM has the following advantages over linear regression:		
(a) Gives faster results - shorter computing time.	□	□
(b) More likely to get a significant p-value.	□	□
(c) Can handle multicolinearity.	□	□
(d) Can model the measurement error of independent variables.	□	□
Question 4. The following statements are true?		
(a) SEM is more appropriate for exploratory rather than confirmatory models.	□	□
(b) Latent variables are variables that could not be directly measured from a person.	□	□
(c) SEM could used to model causation and interdependence relationships.	□	□
(d) SEM is the "King" of all analyses - that is, use it for all analyses.	□	□
Question 5. SEM is appropriate for the following studies:		
(a) To determine the relationships between depression and outcomes of chronic illness giving.	□	□
(b) To determine the predictors of stroke from blood tests data.	□	□
(c) The efficacy of a weight-loss therapy on pre-post estimates.	□	□
(d) The impact of family burden on outcome among patients with bipolar disorder.	□	□

Doctor's particulars:

Name in full: ______________________________

MCR number: ______________________ Specialty: ______________________

Email address: ______________________________

Submission instructions:

A. Using this answer form

1. Photocopy this answer form.
2. Indicate your responses by marking the "True" or "False" box ☑
3. Fill in your professional particulars.
4. Post the answer form to the SMJ at 2 College Road, Singapore 169850.

B. Electronic submission

1. Log on at the SMJ website: URL <http://www.sma.org.sg/cme/smj> and select the appropriate set of questions.
2. Select your answers and provide your name, email address and MCR number. Click on "Submit answers" to submit.

Deadline for submission: (December 2005 SMJ 3B CME programme): 12 noon, 25 January 2006

Results:

1. Answers will be published in the SMJ February 2006 issue.
2. The MCR numbers of successful candidates will be posted online at <http://www.sma.org.sg/cme/smj> by 20 February 2006.
3. All online submissions will receive an aotomatic email acknowledgment.
3. Passing mark is 60%. No mark will be deducted for incorrect answers.
4. The SMJ editorial office will submit the list of successful candidates to the Singapore Medical Council.