A simple introduction to Structural Equation Modelling using Stata: A psychosocioecological approach

Daniel R. Y. Gan
daniel.gry@u.nus.edu

Introduction

This document gives a simple introduction to structural equation modelling using Stata and was created in October 2019 for the Oxford Institute of Population Ageing (OIPA), University of Oxford. The data "GLAS_ESEA2" is confidential and may be requested from OIPA for research purposes.

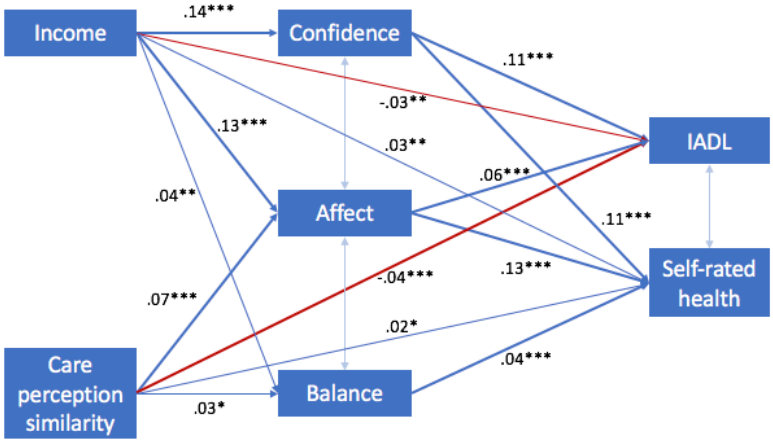
Structural Equation Modelling

Structural equation modelling was developed to support causal inferences. Causal relationships should not be confused with associational or correlational relationships. Causal relationships are directional. For example, the sale of ice-cream and the number of persons drowning during a hot summer may be (spuriously) associated, but they are not causal. At the very least, causality from A to B requires A to occur first, and B to occur only if A occurs. The defining characteristic of structural equation modelling is its statistical use of theoretical assumptions as structural limits during analysis. In other words, the statistical analysis becomes powerful when and only when coupled with theory on their causality.

By default, structural equation modelling assumes that all variables in a causal model are related. As the structural equation modelling is conducted, some of these relationships will be found to be non-existent. This fundamental assumption and approach requires one to carefully select variables of interest and categorise them theoretically into exposure, intervening, and outcome variables.

For example, given a psychosocioecological approach to understanding older population's health, you may theorise that the environment (exposure variable; e.g., household size, household income, care perception similarity) affects individuals' psychosocially (intervening variable; e.g., confidence, affect, balance) which in turn affects their health (outcome variable; e.g., instrumental activities of daily living and self-rated health). Theorised causal direction are typically shown from left to right. As such, exposure, intervening and outcome variables will be arranged to the left, centre and right respectively. After iterating several models, you may find that relationships between some variables are nonsignificant as per in Figure 1. For example, the relationship between care perception similarity and confidence was nonsignificant. Hence the arrow between them was omitted.

Figure 1: Final model taken from Gan, Saxena, Leeson (2019). Single arrows indicate significant causal relationships; vertical double arrows indicate significant non-causal relationships between variables in the same category.



Likelihood ratio	
chi2_ms(10)	7.276
p > chi2	0.699
chi2_bs(56)	9694.773
p > chi2	0.000
Population error	
RMSEA	0.000
90% CI, lower bound	0.000
upper bound	0.011
pclose	1.000
Information criteria	
AIC	199652.194
BIC	200048.561
Baseline comparison	
CFI	1.000
TLI	1.002
Size of residuals	
SRMR	0.004
CD	0.613

For structural equation modelling to be meaningful, at least four variables of interest would be included in addition to co-variables (e.g., sociodemographic). Generally, the sample size to model parameter (i.e., the sum of the number of dyads and the number of variables) ratio should be more than 5:1 to allow convergence. More variables will require greater sample size. Structural equation modelling is generally not recommended for sample sizes of less than 200. For causal inference, a theoretically-informed initial model is paramount regardless of sample size.

Modelling

In the initial model, all variables of interest (i.e., excluding co-variables) should be made to relate. The number of dyad (i.e., relationship between two variables) is dependent on the number of variables. Four variables will give 4x3=12 dyads; five will give 5x4=20 dyads; six will give 6x5=30 dyads etc. Note the exponential increase of dyads with more variables. These, together with variable-co-variable dyads, will contribute to the degrees of freedom available in a model.

For co-variables, they are made to relate with variables of interest if and only if they are correlated at the zeroth order, i.e., significant correlations are found in pairwise regression analysis. For example, one may consider age, education and gender as co-variables in addition to the above variables of interest. Given results of pairwise regression analyses may be per Table 1, age and education will be made to contribute to income, confidence, affect, IADL and self-rated health, whereas gender will contribute to all variables of interest. Co-variables are typically not shown in the final model. That said, information summarised in the above table should be made available to readers.

Table 1: Summary of significant co-variables at zeroth-order

Significant co-variables	**inc**	**cps**	**conf**	**aff**	**bal**	**iadl**	**srh**
Age	Yes		Yes	Yes		Yes	Yes

Education	Yes		Yes	Yes		Yes	Yes
Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes

After setting out and running the initial model, we would remove nonsignificant relationships iteratively until no theoretically acceptable modifications are recommended by Stata to obtain the final model. That said, the structural relationships in Table 1 should be retained throughout the modelling process. They should not be omitted even if found to be nonsignificant during SEM. If the model does not converge, the model may be theoretically problematic or there may be issues with the degrees of freedom. Reducing the number of variables in the initial model by may help.

Recall:
By convention, significance or p-value (P>|t| in Stata) of

- 0.001 or less are deemed extremely significant (***);
- 0.010 or less deemed very significant (**);
- 0.050 or less deemed significant (*);
- 0.100 or less deemed approaching significance (†).

A relationship between two variables is said to be significant with p-value of 0.050 or less.

Goodness of fit

The final model is assessed for its goodness of fit. The causal theory should be reconsidered if good model fit is not found. Per convention, comparative fit index (CFI) and Tucker-Lewis index (TLI) should be above acceptable value of 0.90. Root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) should be below acceptable threshold of 0.10. Chi-squared should be nonsignificant for good model fit. Visualising the level of significance and identifying inverse relationships may aid interpretations of the results as shown in Figure 1. Figure 1 was drawn using Powerpoint for control over line weights. One may also visualise the final model using Stata.

Getting started

Copy the code for each step into Stata's command box, modifying as necessary, and press enter.

1. Import into Stata an Excel file with first row containing variable names. For Mac, use /Users/yourname/Desktop/GLAS_ESEA2.xlsx or .dta without quotations.

```
import excel D:\GLAS_ESEA2.xlsx, firstrow clear
save "D:\GLAS_ESEA2.dta"
use "D:\GLAS_ESEA2.dta", clear
```

2. Check the co-variables for significance at zeroth-order. You should get findings per Table 1.

```
regress hhs age, beta
regress hhs edu, beta
regress hhs fem, beta
```

```
regress inc age, beta
regress inc edu, beta
regress inc fem, beta
regress conf age, beta
...
```

3. Formulate the initial model based on a causal theory. In the initial model, each variable is assumed to contribute to all variables to its right. Include co-variables (highlighted) per findings from Step 2. After the comma, allow error terms of the variables of interest in the same category to co-vary. You'll need to remove the line breaks if you're copying the commands below from a pdf file.

```
sem(hhs->conf aff bal iadl srh) (inc->conf aff bal iadl srh)
(cps->conf aff bal iadl srh) (conf->iadl srh) (aff->iadl srh)
(bal->iadl srh) (age->inc conf aff iadl srh) (edu->inc conf aff iadl
srh) (fem->inc cps conf aff bal iadl srh), stand cov(e.hhs*e.inc)
cov(e.inc*e.cps) cov(e.hhs*e.cps) cov(e.conf*e.aff) cov(e.aff*e.bal)
cov(e.conf*e.bal) cov(e.iadl*e.srh)
```

4. Remove nonsignificant relationships between variables of interest (and only variables of interest) iteratively by modifying the above command. For example, if the relationship between care perception similarity and confidence is nonsignificant, remove 'conf' from the above command. Run the command again. If the co-variance between error terms of confidence and balance are nonsignificant, remove 'cov(e.conf*e.bal).' Continue removing nonsignificant relationships between variables of interest until all remaining relationships between variables of interest are significant. Do not alter the highlighted portion at any time even if the relationships are nonsignificant in SEM.

```
sem(hhs-> conf aff bal iadl srh) (inc->conf aff bal iadl srh)
(cps->~~conf~~ aff bal iadl srh) (conf->iadl srh) (aff->iadl srh)
(bal->iadl srh) (age->inc conf aff iadl srh) (edu->inc conf aff iadl
srh) (fem->inc cps conf aff bal iadl srh), stand cov(e.iadl*e.srh)
cov(e.conf*e.aff) cov(e.aff*e.bal) cov(e.conf*e.bal)
cov(e.hhs*e.inc) cov(e.inc*e.cps) cov(e.hhs*e.cps)

sem(hhs-> conf aff bal iadl srh) (inc->conf aff bal iadl srh) (cps->
aff bal iadl srh) (conf->iadl srh) (aff->iadl srh) (bal->iadl srh)
(age->inc conf aff iadl srh) (edu->inc conf aff iadl srh) (fem->inc
cps conf aff bal iadl srh), stand cov(e.iadl*e.srh)
cov(e.conf*e.aff) cov(e.aff*e.bal) ~~cov(e.conf*e.bal)~~
cov(e.hhs*e.inc) cov(e.inc*e.cps) cov(e.hhs*e.cps)
```

5. Use the modification indices command to check for Stata's recommendations. Only follow theoretically acceptable recommendations, i.e., ignore recommendations with causal direction from left to right. Otherwise, reconsider your causal theory. Repeat Step 4 and 5 until no theoretically acceptable recommendations are found.

```
estat mindices
```

6. Check and screenshot the goodness of fit using the command below.

```
estat gof, stats(all)
```

7. Draw the final model indicating the coefficients and significance of all relationships between variables of interest in Stata or Powerpoint etc. Omit nonsignificant relationships.

References

- Bollen, K. A., & Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research*. Springer.
- Gan, D. R. Y., Saxena, A., & Leeson, G. W. (2019). Changing care perceptions and sense of self in urban East-Southeast Asia? Unpublished.
- Gan, D. R. Y. (2019). *Pathways between neighbourhood experiences and mental health amongst community-dwelling older adults: Towards an urban community gerontology*. PhD dissertation.
- Geiser, C. (2013). *Data analysis with Mplus*. New York: Guilford Press.
- Pearl, J. (1998). Graphs, Causality, and Structural Equation Models. *Sociological Methods & Research*, *27*(2), 487–534. https://doi.org/10.1177/0097700486012004O3
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
- Pearl, J. (2012). The Causal Foundations of Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 68-91). New York: Guilford Press.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson.

Please cite as: Gan, D. R. Y. (2019, October 29). A simple introduction to Structural Equation Modelling using Stata: A psychosocioecological approach. Retrieved from https://blog.nus.edu.sg/healthyageinginplace/data/