Open Peer Review on Qeios

Structural equation modelling

Arindam Basu[1]

1 University of Canterbury

A structural equation model can be thought of a combination of confirmatory factor analysis and linear regression. In structural equation model, the analyst starts with a variance-covariance matrix of variables and specify relationships among the variables. The model is then assessed or estimated with the data provided so that from the data matrix a model matrix is computed. The model and the data matrices are then compared to identify the match between the two matrices. If the two matrices converge, then a solution is arrived at. This solution can be re-examined or modified by the analyst to arrive at the best fit model.

Structural equation model therefore has four steps:

1. Identification - where the analyst identifies the number of free parameters to be estimated from the data provided. If the number of unknown parameters to be estimated exceed the amount of non-redundant data provided in the variance-covariance matrix, the model would be referred to as underidentified. An underidentified model cannot converge to a solution. If the amount of non-redundant information exactly matches the number of unknown parameters to be estimated, the model is said to be just-identified. If the amount of non-redundant information exceeds the number of unknown parameters to be estimated, such a model is referred to as over-identified model. Only over-identified and just-identified model are estimable. For a model with n variables, the amount of non-redundant information is given by the formula: n * (n + 1) / 2; thus, if a model has 4 measured variables, then the amount of non-redundant information in the variance-covariance matrix is 4 * 5 /2 = 10. With this information, the structural equation model solution will only converge if only 10 or fewer parameters are assessed.
2. Model specification. -- This step follows model identification. In the step of model specification, the analyst specifies the relationships between different entities in the model. All variables that are measured are referred to as manifest variables. All variables that are unobserved and inferred from the manifest variables are referred to as latent variables. The variables can be connected to each other using directed paths. A straight directed path can connect a variable A with another variable B. The variable from where a path originates is referred to as exogenous variable and the variable

that receives the path (where the head of the path is directed) is referred to as endogenous variable. All exogenous variables have variances, while all endogenous variables do not have separate variances, but they have error variances as other latent variables. In path analysis notations, straight arrows indicate paths that connect two variables; curved arrows indicate variances or covariances depending on the origin and destination of the curved arrows: if a curved arrow connects two separate variables then such a bidirectional arrow is referred to as covariance; on the other hand if an arrow begins and ends in the same variable, such a curved bidirectional arrow is referred to as variance. Path coefficients are values assigned to each path.

3. Model estimation. -- In this step, the model parameters are estimated and fit statistics are computed. The parameters are path coefficients (beta coefficients for regression models and variances and covariances). The fit statistics are essentially two classes of statistics - based on the chi-square tests of the closenes of fit, where the null hypothesis is that, a high p-value would indicate closeness of fit. The degree of freedom = total number of non-redundant information - total number of free parameters that need to be measured. For example, if a model were to include four variables (thus 10 non-redundant information in the variance-covariance matrix), and eight paths to be estimated, then the degree of freedom would be two (10 - 8 = 2). The other class of fit statistics are those where the 'closeness of the fit' is presented in the form of a 'low value' of the statistic. For example, the root mean square error of approximation (RMSEA) reported by all structural equation modelling programmes, with values less than 0.09 suggest good fit. These values are also reported with an accompanying 90% confidence interval. If the 90% confidence interval straddles 0.10, then the fit is ambiguous.
4. Model modification. -- In this step, the analyst adjust the model specification to obtain the best fit both in terms of the theory being tested as well as the model parameter estimates and fit statistics obtained. Using a path analysis model, the 'arrows' between variables are either deleted or added, or new covariances are specified;

Software programmes used for structural equation modelling

Lisrel - A windows only programme for structural equation modelling and was originally written by Karl Joreskog [1]

Amos - A software bundled with SPSS[2]

Stata - Part of stata statistical analysis software[3]

R - SEM package from the personality project [4]

R - Lavaan - a free and open source structural equation modelling software [5]

R - OpenMx - a free and open source structural equation modelling package in R[6]

Ωnyx - a graphical package for drawing path diagrams and structural equation modelling [7]

Semopy - a python package that has similar syntax as lavaan [8]

References

1. ^ *http://www.ssicentral.com/index.php/products/lisrel/*
2. ^ *https://www.ibm.com/us-en/marketplace/structural-equation-modeling-sem*
3. ^ *https://www.stata.com/features/structural-equation-modeling/*
4. ^ *http://personality-project.org/r/r.sem.html*
5. ^ *http://lavaan.ugent.be/*
6. ^ *https://openmx.ssri.psu.edu/*
7. ^ *http://onyx.brandmaier.de/*
8. ^ *https://arxiv.org/abs/1905.09376*